Generalised Linear Mixed Models: Likelihood and Bayesian Computations with Applications in Epidemiology

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Rafael Sauter

von

Sulgen TG

Promotionskomitee

Prof. Dr. Leonhard Held (Vorsitz)
Prof. Dr. Christel Faes
Prof. Dr. Reinhard Furrer
Prof. Dr. Huldrych Günthard
Prof. Dr. Torsten Hothorn

Zürich, 2015

Preface

The last three and a half years remind me of a journey, which had an entirely unknown destination at its start. The good thing was, even though I had to walk on my own, I rarely walked alone. I would like to thank several people who supported and accompanied me.

First I would like to thank Leonhard Held who made it possible for me to venture this journey and prevented it from becoming an odyssey. I am grateful for his enduring support, his patience and his close mentoring during all these years.

I would also like to thank the members of my dissertation committee, especially Christel Faes for writing the report. Further I would like to thank the research council of the Swiss HIV cohort study, which provided financial support for the corresponding project and especially to Huldrych Günthard, Bruno Ledergerber and Ruizhu Huang who participated in this project and were always taking part in valuable discussions of the intermediate results.

I would also like to thank Daniel Sabanés Bové and Sebastian Meyer for sharing the office and the everyday life with me. Then I want to specially thank Andrea Riebler who essentially taught me everything I know about INLA and always readily provided support whenever I came to a dead end.

My thanks go also to all my former and present colleagues and members of the Biostatistics Department at this institute, which are without any particular order: Michaela Paul, Julia Braun, Sarah Haile, Andrea Riebler, Daniel Sabanés Bové, Andrea Krause, Sebastian Meyer, Wei Wei, Lorenzo Tanadini, Stefanie Muff, Burkhardt Seifert, Małgorzata Roos, Alois Tschopp, Torsten Hothorn, Heidi Seibold, Manuela Ott, Isaac Gravestok, Rachel Heyard, Eva Furrer, Sinikka Kohler, Evelyn Bielser, Niels Hagenbuch and Beate Sick. This list is certainly non-exhaustive and could be extended by many others as there are numerous other enriching people at the EBPI of the University of Zurich who contributed to make my time here enjoyable.

Finally I would like to thank again Małgorzata Roos, Andrea Riebler, Julia Braun and Sebastian Meyer for reading and correcting the introduction chapter of this thesis.

Zurich, July 2015

Rafael Sauter

Zusammenfassung

Wiederholtes messen desselben Patienten impliziert, dass die erhobenen Beobachtungen nicht unabhängig sind, da diese von denselben patientenspezifischen Eigenschaften beeinflusst werden. Ein generalisiertes lineares gemischtes Modell (GLMM) berücksichtigt diese Abhängigkeiten, indem patientenspezifische Modellparameter eingeführt werden, die als zufällige Effekte bezeichnet werden. Die Struktur der Abhängigkeiten in den Daten kann Formen annehmen, die verschieden sind von der, welche durch wiederholtes beobachten derselben Patienten auftritt. Es kann eine zeitliche, räumliche oder zeit-räumliche Abhängigkeit, im zugrunde liegenden Prozess, vorhanden sein. Auch ein Netzwerk aus verschiedenen Einheiten, die verbunden sind und wiederholt beobachtet werden, kann den Einschluss von zufälligen Effekten in einem GLMM motivieren.

Ein GLMM schätzt, bei gegebener Struktur der zufälligen Effekte, den bedingten Erwartungswert der interessierenden Parameter, die als fixe Effekte bezeichnet werden. Die Likelihood Inferenz bestimmt die bedingten Schätzwerte durch numerische Integration über die zufälligen Effekte, da dieses Problem generell nicht analytisch lösbar ist. Die numerische Integration kann rechnerisch schwer lösbar sein, je nach Komplexität der Struktur der zufälligen Effekte und der verfügbaren Daten.

Ein Bayesianischer Inferenz Ansatz bildet die Struktur der zufälligen Effekt, unter Einschluss von Priori-Verteilungen für diese Parameter, ab. Der Einschluss von Priori-Verteilungen ist flexibel und kann die unterschiedliche, verfügbare Information auf verschiedenen Ebenen des Modells abbilden. Bayesianische Inferenz wird üblicherweise mit einer *Markov chain Monte Carlo* (MCMC) Simulation durchgeführt, die eine grosse Rechenleistung verlangt. Falls der Struktur der zufälligen Effekte ausschliesslich Gaussche Priori-Verteilungen zugewiesen werden, nur eine zusätzliche Ebene von Hyperparametern und eine beschränkte Ordnung der Abhängigkeiten zwischen den Einheiten angenommen wird – so dass ein Gaussches Markov Zufallsfeld resultiert – kann die Methode der *integrated nested Laplace approximations* (INLA) als Alternative zu MCMC verwendet werden. INLA verlangt weniger Rechenleistung, was insbesondere für komplexe Modelle ein Vorteil ist.

Diese Dissertation untersucht beide Inferenz Methoden für GLMMs, diskutiert damit verbundene rechen-technische Aspekte und erläutert diese anhand mehrerer epidemiologischen Anwendungen. Als Erstes wird die Likelihood Inferenz für ein linear gemischtes Modell, basierend auf longitudinale Daten aus der Schweizerischen HIV Kohortenstudie durchgeführt. Das Modell untersucht, ob vorherig beobachtete Lymphozyt-Subtypen relevante Prädiktoren für den Krankheitsverlauf von unbehandelten und behandelten HIV infizierte Patienten sind. Im darauf folgenden Teil wird diskutiert wie die spezielle Situation, bei welcher patientenspezifische longitudinale Profile keine Variation in der Ausgangsgrösse haben, die Likelihood und Bayesianische Inferenz mit INLA beeinflussen. Wir zeigen, dass mit einem zunehmenden Anteil an Patienten, welche keine Variation in der Ausgangsgrösse haben, die Maximum likelihood (ML) Schätzung der Parameter, in einem Modell mit einer binären Ausgangsgrösse, numerische Probleme verursacht. Weiterhin zeigen wir, dass in einem solchen Fall INLA Schätzungen generiert, die weder mit ML noch mit MCMC Schätzungen übereinstimmen. Im dritten Teil diskutieren wir wie die besondere Abhängigkeitsstruktur einer Netzwerk Meta-Analyse, unter Berücksichtigung der versuchsspezifischen Heterogenität und möglicher Inkonsistenzen im Netzwerk, mit INLA implementiert wird. Der letzte Teil der Dissertation untersucht die Verwendung von informativen Priori-Verteilungen, welche adaptive Gewichte verwenden, die anhand der beobachteten Daten bestimmt werden. Üblicherweise werden nicht informative und unkorrelierte Priori-Verteilung für die fixen Effekte in einem GLMM angenommen. In manchen Situationen kann diese Annahme zu unrealistischen Parameter Schätzungen führen. Adaptives gewichten der Priori-Verteilungen, basierend auf den beobachteten Daten und unter Einschluss von Korrelationen, kann dazu dienen dieses Problem zu beheben.

Abstract

Repeatedly observing the same patient implies that these samples will not be independent, as they are affected by the same common patient-specific characteristics. A generalized linear mixed model (GLMM) takes this dependency structure into account by introducing patient-specific model parameters which are called random effects. The dependency structure in the collected data could have various forms, though other than the one which arises from repeatedly observing patients in a study population. A temporal, spatial or even spatio-temporal pattern may be present in the underlying sampling process. Or a network of different clusters which are connected and repeatedly observed may motivate the inclusion of random effects in a GLMM.

Given the random effect structure, a GLMM investigates the conditional expectation for the parameters of interest, which are called fixed effects. In likelihood inference, the conditional estimates are determined by numerically integrating over the random effects, as in general this problem is not analytically solvable. The numerical integration may be computationally difficult to solve, depending on the complexity of the random effect structure and the data at hand.

A Bayesian inference approach maps the random effect structure by including prior distributions for these parameters. The inclusion of prior distributions is flexible and may reflect different stages of information at different levels of the model. Bayesian inference is commonly carried out using computationally intensive Markov chain Monte Carlo (MCMC) sampling. If exclusively Gaussian priors are assigned to the random effect structure, with only one additional level of hyperparameters and a limited order of dependencies between clusters – such that a Gaussian Markov random field results – one can apply integrated nested Laplace approximations (INLA). INLA is an alternative to MCMC and requires less computational effort, which especially for complex models is an huge advantage.

This thesis investigates both inference approaches for GLMMs, discusses related computational issues and illustrates these with several epidemiological applications. First, likelihood inference is carried out for a model based on longitudinal data from the Swiss HIV cohort study. This model investigates if past lymphocyte subtypes are relevant predictors for the disease progression among untreated and treated HIV infected patients. In the second part we discuss how the special situation, in which patient-specific longitudinal profiles show no variation in the response, influences the likelihood and Bayesian inference with INLA. We show that, with an increasing proportion of patients who have no variation in the response, numerical issues arise in the Maximum likelihood (ML) estimation of a binary response GLMM. Furthermore, we show that in this case INLA produces estimates that are inconsistent with ML or MCMC inference. In the third part we discuss how the particular dependency structure of a network meta-analysis is implemented with INLA, taking into account trial specific heterogeneity and possible network inconsistencies. The last part of the thesis examines the use of informative priors which use adaptive weights that are based on the observed data. Usually the prior distributions for the fixed effects in a GLMM are assumed to be uninformative and uncorrelated. In some situations this assumption may lead to unrealistic parameter estimates. An adaptively weighted informative prior distribution may help to resolve this problem.

Thesis outline

Introduction

Paper I: **CD8 counts and CD4/CD8 ratio independently predict CD4 response in drug naive and in patients on cART**

Rafael Sauter, Ruizhu Huang, Bruno Ledergerber, Manuel Battegay, Enos Bernasconi, Matthias Cavassini, Hansjakob Furrer, Matthias Hoffmann, Mathieu Rougemont, Huldrych F. Günthard, Leonhard Held & the Swiss HIV cohort study.

Paper submitted to *Journal of Acquired Immune Deficiency Syndromes*.

Paper II: **Quasi-complete Separation in Random Effects of Binary Response Mixed Models: Integrated Nested Laplace Approximations vs. MCMC**

Rafael Sauter, Leonhard Held.

Paper published in *Journal of Statistical Computation and Simulation*.

Paper III: **Network meta-analysis with integrated nested Laplace approximations**

Rafael Sauter, Leonhard Held.

Paper published in *Biometrical Journal*.

Paper IV: **Adaptive prior weighting in generalized linear models**

Leonhard Held, Rafael Sauter.

Paper in revision for *Biometrics*.

Introduction

Statistical models describe deterministic and stochastic components of a data generating process by using as few parameters as necessary. In most applications model parameters map an underlying structure related to the problem. Some information about this structure may be known and thus may serve to adequately incorporate dependencies between model parameters. Repeatedly observing the same entities (*e. g.* patients), or collecting several observations under different environments (*e. g.* hospitals), will inherently induce possible dependencies within these entities or circumstances, which differ from the ones between entities. Taking into account and parameterizing such entity-specific dependencies is necessary, such that the observations can be considered to be conditionally independent, given the entity-specific parameters. This conditional independence implies the exchangeability of the observed entities which is crucial if one considers to carry out inference for the model parameters.

The recognition and formalisation of the statistical analysis for such repeated and dependent observations dates back to more than one hundred years. Models developed during these days and for a long time thereafter were limited in their applications. Progress in the development of statistical methods but also the increasing availability of computers and more and more computing power lead to the dissemination of such models for repeated measurements. A long history of developing methods suited for particular applications, such as non-normal distributed outcomes and trials with unequally observed or unbalanced data, finally lead to the comprehensive generalized linear mixed model (GLMM) framework: a regression model class for outcomes with a distribution from the exponential family and with different types of entity-specific effects.

The generic concept of GLMMs allows to address a broad set of different applications: the dependency structure may come from independent entities, from a spatial, temporal, or combined spatio-temporal pattern, or may describe dependencies in any other connected graph. Longitudinal data for epidemiological studies are one of the most prominent examples, but also a meta-analysis which describes repeated observations of the same treatment comparison may make use of GLMMs. The success of GLMMs during the last years was supported by an increasing number of sophisticated, ready to use software. However, implemented algorithms sometimes put limitations on the distribution or the dependencies between and within different entities. Such limitations are mainly driven by the complexity of the problem, *e. g.* numerical integration. Computing issues arise in Bayesian as well as in likelihood based inference for GLMMs. The strength of one inference approach may be the others weakness. Either way it is important to recognize similarities and limitations in both practices.

This chapter introduces GLMMs, including likelihood and Bayesian inference and is structured in the following way: Section 1 starts with a short historical outline of the origins and the milestones in the development of the GLMM framework. Subsequently the model assumptions and data structure for GLMMs will be introduced and extensions as well as related model classes will be discussed shortly. In Section 2 the focus is on likelihood inference, followed by a introduction of the Bayesian inference approach to GLMMs in Section 3. Sections 2 and 3 will be complemented with a short discussion of limitations and by a description of available software for each inference approach.

1 Generalized linear mixed models

1.1 The origins

According to Scheffé (1956) the idea of including different error terms for observations collected under different circumstances was probably first formally noted by Airy (1861, Part IV), a British astronomer. He was interested in collecting several observations of the same phenomena with a telescope at several nights. It was this setup which lead him to introduce in his model a special variance component for each night, reflecting the specific but varying circumstances, *e. g*. in the atmosphere or the personal condition, encountered during each night. With this work he described the foundations of a linear mixed model, preceding the work by Fisher (1918, 1925) who laid out the same problem in a more formal way in the context of the analysis of variance (ANOVA).

The univariate repeated measures ANOVA is a precursor of the linear mixed model (LMM) which is applicable to outcomes with a normal distribution. In an ANOVA, covariates must be discrete factors, observations must be balanced, the covariance structure among repeated measures is restricted and variances assumed to be constant. The regression approach of LMMs relaxes these assumptions and allows for unbalanced and unequally spaced data, such that the number of observations per entity need not be the same and time periods between subsequent observations can vary. The inclusion of continuous covariates and more general correlation structures for the variance within an entity is also possible with a LMM. Contributions which were relevant in establishing LMMs also came from the interest to analyse growth curves, for which individual-specific random effects were introduced and regression coefficients were allowed to vary across individuals (Wishart, 1938; Rao, 1958). The basic principles for LMMs were established early and fundamental work has been successively added *e. g*. by Harville (1976, 1977). It took some time until the application of LMMs to longitudinal data was discussed by Laird and Ware (1982).

Naturally, also models with random effects for discrete outcomes were of interest and there are several contributions, *e. g*. for binary data (Ashford and Sowden, 1970; Cox, 1972). In parallel Nelder and Wedderburn (1972) and McCullagh and Nelder (1989) established the generalized linear model (GLM), a framework for regression models with outcomes from any distribution that is part of the exponential family. The combination of the LMM and of GLMs lead to the extension of the model framework for repeated measures. The two methodological strains were formally embraced by the GLMM definition, described and illustrated with a broad variety of applications by Breslow and Clayton (1993). A GLMM is a regression model for an outcome from the exponential distribution family which takes dependencies for repeated observations from the same entities into account by introducing random effects. The term GLMM appears to be introduced already by Gilmour *et al*. (1985). Ideas for applying Bayesian inference to GLMMs evolved in parallel and were discussed by Karim and Zeger (1992).

The mixed effects model approach must be distinguished from a second model framework, which is also concerned about dependencies arising from repeated measures. This other model class, known under the term marginal model, describes the population mean. This is in contrast to mixed models, also called random effects models, which investigate the entity-specific mean, conditional on the entity-specific parameters. In contrast to mixed models, marginal models do not specify the full distribution but only make assumptions for the first two moments. The estimation of marginal models with generalized estimating equations (GEE) was proposed by Liang and Zeger (1986) and is a generalization of the quasi-likelihood approach by Wedderburn (1974). An overview of the development of GLMMs with a focus on longitudinal data can be found in Fitzmaurice, M., Verbeke and Molenberghs (2008).

1.2 Model and data structure

A good introduction to GLMMs can be found in Fahrmeir *et al.* (2013, Chapter 7). The outcome of interest y_{ij} is repeatedly observed for entities $i = 1, \ldots, m$ at occasions $j = 1, \ldots, n_i$. The number of observation per entity n_i does not have to be equal, thus a GLMM can handle unbalanced designs and the total number of observations is $N = \sum_{i=1}^{m} n_i$. The distribution of y_{ij} comes from the exponential family such that the GLM framework (McCullagh and Nelder, 1989) can be applied. In a GLMM the conditional expectation $E(y_{ij} \mid \mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i)$ is linked to a linear predictor η_{ij} with a monotone link function $h(\cdot)^{-1}$

$$h^{-1}\{E(y_{ij} \mid \mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i)\} = \eta_{ij} = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{z}_{ij}^\top \mathbf{b}_i \tag{1}$$

where $\mathbf{x}_{ij}$ is a vector of covariates of length p, including an intercept and $\boldsymbol{\beta}$ a vector of the same length with the parameters of interest, also called fixed effects. Usually $\mathbf{z}_{ij}$ is a sub-vector of $\mathbf{x}_{ij}$ of length $q < p$ and $\mathbf{b}_i$ is a vector with entity-specific parameters, or random effects, of length q. In a random intercept model q is equal to one and $z_{ij} = 1$. The linear predictor in Equation (1) includes one level of random effects $\mathbf{b}_i$. Of course one could imagine that there exist several different or nested levels of entities for which random effects could be required to correctly reflect dependencies. For the sake of keeping the notation simple we here restrict the model to one level of random effects only. Possible extensions are discussed in Section 1.3.

Aggregating the data at each entity-specific level illustrates how the design matrices must be organised. All observations of one cluster i are contained in the vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{ij}, \ldots, y_{in_i})^\top$ of length n_i. Then the linear predictor for cluster i is

$$\boldsymbol{\eta}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i \tag{2}$$

where now

$$\mathbf{X}_i = \begin{pmatrix} \mathbf{x}_{i1}^\top \\ \vdots \\ \mathbf{x}_{ij}^\top \\ \vdots \\ \mathbf{x}_{in_i}^\top \end{pmatrix}, \quad \mathbf{Z}_i = \begin{pmatrix} \mathbf{z}_{i1}^\top \\ \vdots \\ \mathbf{z}_{ij}^\top \\ \vdots \\ \mathbf{z}_{in_i}^\top \end{pmatrix}$$

and $\mathbf{X}_i$ is a fixed effects design matrix of dimension $n_i \times p$ and $\mathbf{Z}_i$ a random effects design matrix of dimension $n_i \times q$. Aggregating the data to the next level, across all entities, results in a vector with all observations $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_i, \ldots, \mathbf{y}_m)^\top$ of length N, such that the linear predictor is

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}.$$

The fixed effects design matrix $\mathbf{X}$ of dimension $N \times p$ and the random effects design matrix $\mathbf{Z}$ of dimension $N \times qm$ are defined as

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1^\top \\ \vdots \\ \mathbf{X}_i^\top \\ \vdots \\ \mathbf{X}_m^\top \end{pmatrix}, \quad \mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 & & & \mathbf{0} \\ & \ddots & & \\ & & \mathbf{Z}_i & \\ & & & \ddots \\ \mathbf{0} & & & \mathbf{Z}_m \end{pmatrix}$$

and each cluster i has q random effects, such that the vector $\mathbf{b}$ has length qm.

The GLMM is complemented by the assumption that the entity-specific random effects $\mathbf{b}_1, \ldots, \mathbf{b}_m$ are independent and follow the same multivariate normal distribution

$$\mathbf{b}_i \sim \mathrm{N}(\mathbf{0}, \mathbf{D}),$$

where $\mathbf{D}$ is a $q \times q$ covariance matrix. The zero expectation of $\mathbf{b}_i$ implies that random effects are symmetric deviations from the respective population mean, which is an element of $\mathbf{X}\boldsymbol{\beta}$. As random effects between entities are uncorrelated, it follows that $\mathbf{b} \sim \mathrm{N}(\mathbf{0}, \mathbf{G})$ and $\mathbf{G} = \mathrm{diag}(\mathbf{D}_1, \ldots, \mathbf{D}_m)$, which is a positive definite and block-diagonal covariance matrix of dimension $qm \times qm$.

The following two examples give an impression on how diverse the observed data structure for repeatedly observed entities can be.

Longitudinal data

A specific disease in a study population is often observed repeatedly at several occasions for the same patients. Such longitudinal data has two sources of possible dependencies: one is from repeatedly observing the same patient, the other from possible temporal dependencies for pairs of observations from the same patient *i. e.* serial correlation.

Based on the notation introduced above, i would be a subscript identifying one among m different patients that was observed at occasion j. The observations are ordered by the times t_{ij} at which they took place, which defines a sequence $(t_{i1}, \ldots, t_{ij}, \ldots, t_{in_i})$. Usually the temporal ordering is considered to imply a causal relationship as described by Diggle (2002, Chapter 12). If besides a random intercept one also assumes a serial correlation between observations, then the linear predictor in Equation (2) is supplemented by an additional term such that

$$\boldsymbol{\eta}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{W}_i(t_{ij}), \tag{3}$$

where $\mathbf{W}_i(t_{ij})$ are independent realizations from a stationary Gaussian process with mean zero, variance ν^2 and correlation function $\rho(\,|\,t_{ij} - t_{ik}\,|\,)$ (see Chapter 5 in Diggle, 2002). The correlation function captures the serial correlation of the stochastic process. This functional relationship can be defined as *e. g.* an exponentially decaying correlation function or an autoregressive process for discrete, equally spaced observations.

The observed outcomes for the longitudinal observations may follow a normal distribution, which requests a LMM, such as for the square root transformed CD4 lymphocyte counts in HIV-1 infected patients presented in Paper I. The outcome may also be from any other distribution of the exponential family, which implies a GLMM, such as the Bernoulli distributed data for the probability of having a toenail infection presented in Paper II. The analysis of longitudinal data with GLMMs is well described in Diggle (2002), by Verbeke and Molenberghs (2000) for LMMs or in Molenberghs and Verbeke (2005) for discrete outcomes.

Network meta-analysis

Evidence for a relative effect of two treatments is usually collected in a series of independent trials. Such a comparison may be extended to a set of different treatments or interventions, which forms a network of treatment comparisons, as discussed in Paper III. A GLMM for such a network meta-analysis usually introduces random effects for comparisons between two treatments, if they were repeatedly observed in different trials. This is motivated by the

assumption of possible heterogeneity in the circumstances, *e. g*. the different study populations which were used in the trials. The trials $i = 1, \ldots, m$ investigate the same relative treatment effect j among the set of different treatments $1, \ldots, T$. The outcome of the GLMM may be the log odds ratio for the relative treatment comparison, which follows a normal distribution (Lumley, 2002). Alternatively, the outcome may also be the number of observed events among all study participants for each trial, which implies a binomial distributed outcome and the relative treatment effect is included as a model parameter (Lu and Ades, 2006).

1.3 Related model classes and extensions

Sometimes it is difficult to find a suitable functional relationship between an observed metric covariate u_{ij} and the outcome y_{ij}. An extension of the linear predictor with a flexible function $f(u_{ij})$ may be adequate. Possible smooth, non-linear functions for L different metric covariates $u_{ij1}, \ldots, u_{ijL}$ can be used as additive terms to extend the linear predictor in Equation (2) to

$$\boldsymbol{\eta}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + f_1(u_{i1}) + \ldots + f_l(u_{il}) \ldots + f_L(u_{iL}). \tag{4}$$

In Equation (3) a similar extension was anticipated by introducing the term $\mathbf{W}_i(t_{ij})$ to capture serial correlation between observations. However, the function may also describe a smooth, non-parametric relationship which leads to a generalized additive mixed model (GAMM) (Ruppert *et al.*, 2003, 2009). Generalized additive models (GAM) without entity-specific random effects but with a linear predictor which combines the component $\mathbf{X}\boldsymbol{\beta}$ with non-parametric functional relationships are discussed by Hastie and Tibshirani (1990). The functional term $f_l(u_{il})$ could also describe a spatial effect for location variables u_{il}, which leads to a geo-additive model. Fahrmeir *et al.* (2004) coin the term structural-additive regression (STAR) for describing a model class which extends the GAMM framework. They relax the additive form of the functional relationships and allow also for non-linear interactions between two metric covariates $f(u_{il}, u_{ik})$ and for functions which have varying effects depending on components of $\mathbf{X}$. In this generic STAR setup the functional relationships can describe entity-specific random effects, a spatial, temporal or combined spatio-temporal structure but also any other form of non-linear dependency which is added to the linear predictor.

The functions $f(u_{il})$ usually depend on a continuous or discrete criterion, such as the distance between two locations or two points in time. For discrete observations in time *e. g*. a random walk can serve as smooth function. In general one can define Markov random fields to introduce conditional dependencies of some order for neighbouring entities, *e. g*. for different regions in space (Rue and Held, 2005). Rue *et al.* (2009) use the term *latent Gaussian model* to define a subgroup of STAR models which uses a Gaussian prior for the components $\boldsymbol{\beta}, \mathbf{b}_i$ and on each function $f(u_{il})$. The Gaussian distribution assumption for the model components is particularly attractive to use in combination with Gaussian Markov random fields (GMRF), as the sparsity of the implied structure has attractive computational properties as described by Rue and Held (2005, Chapter 2).

Here it is appropriate to draw a line to hierarchical models. Hierarchical, or multilevel models are motivated from a Bayesian perspective as discussed by *e. g*. Gelman *et al.* (2014). The hierarchical approach distinguishes different levels of observational units for which information is available. The units or entities on each level in a hierarchical model are exchangeable and each level is described by different model parameters. Hierarchical models can be seen as a broader approach which includes GLMMs as special case, sometimes under the term hierarchical linear models (Gelman *et al.*, 2014, Chapter 5) or hierarchical regression models (Blangiardo and Cameletti, 2015, Chapter 5). Also latent Gaussian fields (Rue *et al.*, 2009),

sometimes called hierarchical GMRFs (Rue and Held, 2005, Chapter 4) fit into the context of hierarchical modelling. Hierarchical GMRFs use the following constitutive elements: on the first level a distribution assumption from the exponential family for observations y_{ij} is determined. The second level assumes a multivariate Gaussian prior distribution for the model components $\boldsymbol{\beta}, \mathbf{b}_i$ and for all functions $f(u_{il})$ in the form of a GMRF. The parameters, which define the covariance structure of the GMRF, build the third level in the model hierarchy. A prior distribution is assigned again to each of these hyperparameters.

2 Likelihood inference

A compact description of likelihood inference for GLMMs can be found in Fahrmeir and Tutz (2001, Chapter 7) or Fahrmeir *et al.* (2013, Chapter 7). In a GLMM we assume that the outcome $\mathbf{y}_i$ is conditionally independent, such that we can write the conditional density for an entity i as

$$f(\mathbf{y}_i \mid \boldsymbol{\beta}, \mathbf{b}_i) = \prod_{j=1}^{n_i} f(y_{ij} \mid \boldsymbol{\beta}, \mathbf{b}_i)$$

where here $f(\cdot)$ is a density or probability mass function from the exponential family. The marginal density $f(\mathbf{y}_i)$ can be determined by integrating over the random effects in the conditional density

$$f(\mathbf{y}_i) = \int f(\mathbf{y}_i \mid \boldsymbol{\beta}, \mathbf{b}) f(\mathbf{b}_i \mid \mathbf{D}(\boldsymbol{\delta})) d\mathbf{b}_i$$

such that the marginal likelihood for all m entities is defined as

$$L(\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\delta}) = \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} f(y_{ij} \mid \boldsymbol{\beta}, \mathbf{b}_i) f(\mathbf{b}_i \mid \mathbf{D}(\boldsymbol{\delta})) d\mathbf{b}_i \tag{5}$$

where $\boldsymbol{\delta}$ are unknown hyperparameters which determine the distribution of the random effects covariance matrix $\mathbf{D}(\boldsymbol{\delta})$.

2.1 Linear mixed models

For a LMM there exists an analytical solution of the integral contained in Equation (5), which results in the marginal distribution $\mathbf{y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\delta}))$, where $\mathbf{V}(\boldsymbol{\delta}) = \sigma^2 \mathbf{I}_N + \mathbf{Z}\mathbf{G}(\boldsymbol{\delta})\mathbf{Z}^\top$. The residual variance is σ^2 and $\mathbf{I}_N$ is the identity matrix of dimension N. The corresponding distribution for the conditional distribution of the LMM outcome is $\mathbf{y} \mid \mathbf{b} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \sigma^2 \mathbf{I}_N)$. For LMMs with unknown random effects covariance structure the inference problem is still challenging: one needs to find estimates for $\boldsymbol{\beta}, \mathbf{b}_i$ and $\boldsymbol{\delta}$. Maximising the LMM likelihood for $\boldsymbol{\beta}$ with a fixed $\boldsymbol{\delta}$ gives the estimate

$$\tilde{\boldsymbol{\beta}}(\boldsymbol{\delta}) = (\mathbf{X}^\top \mathbf{V}(\boldsymbol{\delta})^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{V}(\boldsymbol{\delta})^{-1} \mathbf{y}$$

which can be derived as best linear unbiased estimator (BLUE) (Harville, 1977). If $\boldsymbol{\delta}$ is known, then the estimates for the random effects

$$\tilde{\mathbf{b}} = \mathbf{G}(\boldsymbol{\delta}) \mathbf{Z}^\top \mathbf{V}(\boldsymbol{\delta})^{-1} (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$$

are estimated best linear unbiased predictors (EBLUP) (Harville, 1976). However, if $\boldsymbol{\delta}$ is unknown one can now use a profile likelihood by plugging in $\tilde{\boldsymbol{\beta}}(\boldsymbol{\delta})$ to determine an estimate for

$\boldsymbol{\delta}$, or one could use the marginal likelihood, integrating over $\boldsymbol{\beta}$

$$\int L(\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\delta}) d\boldsymbol{\beta} \tag{6}$$

to determine the estimates for the hyperparameters $\boldsymbol{\delta}$ with a Fisher-scoring algorithm. The marginal likelihood in Equation (6) can be embedded into the restricted maximum likelihood (REML) approach for linear models. The REML estimation corrects the bias of the Maximum likelihood (ML) covariance parameter estimates (Diggle, 2002, Chapter 4.5). For LMMs the ML estimator based on the profile likelihood ignores the loss of degrees of freedom for estimating the fixed effects, thus is biased and so in general the REML approach should be preferred.

2.2 Generalized linear mixed models

In contrast to LMMs there is no analytical solution for the integral in Equation (5) for GLMMs. A GLMM requires a numerical integration over the q-dimensional vector $\mathbf{b}_i$ in the marginal likelihood. There exist different approaches to solve this task. One could apply a Laplace approximation (see Held and Sabanés Bové, 2014, Appendix C). Laplace's method approximates the integral $\int f(x)dx = \int \exp(\log f(x))dx$ by applying a Taylor series expansion around $x^\star$, which is the mode $x^\star = \text{argmax}_x \log f(x)$. This implies that the first derivative at $x = x^\star$ is zero and

$$\log f(x) \approx \log f(x^\star) + \frac{(x - x^\star)^2}{2} \frac{\partial^2 \log f(x)}{\partial x^2}\bigg|_{x=x^\star}$$

which results in an approximation of the integral with a Gaussian kernel

$$\int f(x)dx \approx f(x^\star) \int \exp\left\{-\frac{(x - x^\star)^2}{2\sigma^{\star^2}}\right\} dx$$

where $\sigma^{\star^2}$ is equal to the inverse, negative second derivative for x evaluated at $x^\star$.

Alternatively one can also approximate the marginal likelihood in Equation (5) with a Gauss-Hermite approximation (see Fitzmaurice *et al.*, 2008, Chapter 4). Instead of using $\mathbf{b}$ it is useful to use a Cholesky decomposition of the random effects covariance $\mathbf{D}(\boldsymbol{\delta})$, such that $\mathbf{b}_i = \mathbf{D}(\boldsymbol{\delta})^{1/2}\mathbf{b}_i^\star$, and an independent standard normal distribution for $\mathbf{b}_i^\star \sim \text{N}(\mathbf{0}, \mathbf{I})$ results. For each random effect $b_{ik}^\star$, among q random effects for entity i, the one dimensional integral can be approximated by

$$\int \prod_{j=1}^{n_i} f(y_{ij} \mid \mathbf{b}_i^\star) f(b_{ik}^\star) db_{ik}^\star \approx \sum_{r=1}^{R} w_r \prod_{j=1}^{n_i} f(y_{ij} \mid a_r, \mathbf{b}_{i,-k}^\star)$$

where $\mathbf{b}_{i,-k}^\star$ is a vector with all standardized random effects for entity i, except the kth one and w_r, a_r are the weights and locations of the Gauss-Hermite quadrature rule of degree $(2R - 1)$ (Stroud and Secrest, 1966). The quadrature points and weights can also be defined adaptively (Pinheiro and Bates, 1995), such that they depend on the cluster-specific mean and variance, which results in an improved approximation (Lesaffre and Spiessens, 2001; Rabe-Hesketh *et al.*, 2002). Increasing the number of quadrature points increases the accuracy of the approximation. With a single quadrature point the Gauss-Hermite approximation is equal to the Laplace approximation. There exists no general applicable rule in how one should determine the number of quadrature points, but robust estimates with respect to a changing number of quadrature points is certainly desirable.

Stiratelli *et al.* (1984) proposed to use a penalized quasi-likelihood approach (PQL) for GLMMs. Breslow and Clayton (1993) motivate the PQL estimation for GLMMs with a Laplace approximation to the marginal likelihood. They state that the likelihood

$$L(\boldsymbol{\beta}, \mathbf{b}; \boldsymbol{\delta}) = f(\mathbf{y} \mid \mathbf{b}, \boldsymbol{\beta}) f(\mathbf{b} \mid \mathbf{D}(\boldsymbol{\delta}))$$

can be rewritten as penalized log-likelihood of the form

$$l(\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\delta}) = l(\boldsymbol{\beta}, \mathbf{b}) - \frac{1}{2} \mathbf{b}^\top \mathbf{G}(\boldsymbol{\delta})^{-1} \mathbf{b} \tag{7}$$

where $l(\boldsymbol{\beta}, \mathbf{b}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} f(y_{ij} \mid \boldsymbol{\beta}, \mathbf{b})$ is the log-likelihood of the implied GLM, and the penalization term $-\mathbf{b}^\top \mathbf{G}(\boldsymbol{\delta})^{-1} \mathbf{b}$ follows from the normal distribution assumption for $f(\mathbf{b})$. The PQL approach uses some starting values for $\boldsymbol{\beta}$, $\mathbf{b}$ and $\boldsymbol{\delta}$ and computes working responses which are used to solve the score functions of the penalized likelihood for $\boldsymbol{\beta}, \mathbf{b}$. In a second step the estimates for $\boldsymbol{\delta}$ are found by numerically solving the restricted likelihood with Fisher scoring by using the estimates $\boldsymbol{\beta}$ and $\mathbf{b}$ from the first step. The two steps are iteratively repeated until the required convergence criteria are met. The score functions are the same as for the LMM but with different weights. Thus this iteratively reweighted least squares (IRLS) algorithm can also be applied to LMMs to improve estimates of $\boldsymbol{\delta}$. Estimation with PQL can yield a substantial bias, especially for binary responses with few observations per patient. Therefore there were efforts in adapting the penalization criterion to establish a bias correction by Breslow and Lin (1995) and Lin and Breslow (1996) which was even taken further by using Laplace approximations based on a Taylor series expansion of higher order by Raudenbush *et al.* (2000).

The second common estimation algorithm for GLMMs also involves two steps: first the estimates for $\mathbf{b}$ are determined by a penalized iteratively reweigthed least squares (P-IRLS) algorithm (Bates and DebRoy, 2004) with $\boldsymbol{\beta}$, $\mathbf{b}$ and $\boldsymbol{\delta}$ fixed at some starting values. The penalized least squares criterion is optimized by iteratively updating estimates for $\mathbf{b}$ and then reweighting the working responses until convergence. In the second step the marginal likelihood is approximated with a Laplace or Gauss-Hermite approximation, given the estimates $\mathbf{b}$ from the first step, which is then maximized for $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$. Both steps are repeated until convergence of the deviance $-2l(\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\delta})$ is reached.

In general, ML inference for GLMMs will neglect any uncertainty coming from the estimation of the random effects $\mathbf{b}$. Furthermore, the penalization term in Equation (7) is proportional to $\mathbf{G}(\boldsymbol{\delta})^{-1}$ which reflects the inverse of the random effect variance. If $\mathbf{G}(\boldsymbol{\delta}) \to \infty$ then the penalization term goes to zero, such that $\mathbf{b}$ will not be treated differently than the fixed effects $\boldsymbol{\beta}$. The penalization also increases for increasing deviations from $E(\mathbf{b}) = \mathbf{0}$. The penalization term has an influence on the estimates for $\mathbf{b}_i$, which are shrunk more towards the overall mean $\mathbf{X}\boldsymbol{\beta}$ with increasing entity-specific random effect variance $\mathbf{D}(\boldsymbol{\delta})$. Also fewer numbers per entity n_i result in stronger shrinkage for the corresponding $\mathbf{b}_i$. For LMMs the EBLUP can be shown to be a weighted average between the population averaged mean response profile and the entity-specific response profile and the weight depends on the relation between the entity-specific within variance $\sigma^2 \mathbf{I}_{n_i}$ and the overall variance $\mathbf{V}(\boldsymbol{\delta})$ (Fitzmaurice *et al.*, 2004, Chapter 8.6).

In contrast to that, the extreme case of $\mathbf{D}(\boldsymbol{\delta}) = \mathbf{0}$, which means that *e. g.* for a GLMM with binary outcome the observations for a specific entity have always the same value, the penalization term goes to infinity such that the ML estimate for $\mathbf{b}_i$ will not be defined any more. The problem of non existent ML estimators in a GLM for such a setting is commonly de-

scribed as complete separation, because one covariate perfectly predicts the outcome, or as quasi-complete separation if the covariate predicts a subset of the outcome (Albert and Anderson, 1984). Firth (1993) suggested a penalized likelihood approach to solve this problem for GLMs. However, for GLMMs the complete separation problem may also be present for the entity-specific effects $\mathbf{b}_i$. Depending on the proportion of clusters which have no variation, this cluster-specific quasi-complete separation may cause numerical instabilities in the marginal likelihood approximation.

2.3 Software

Nowadays there exist several software packages for likelihood inference in GLMMs. The following overview is restricted to software packages in `R` (R Core Team, 2015), although other statistical software has similar routines implemented, like `PROC NLMIXED` in SAS or `xtmelogit` for logistic GLMMs in Stata. For `R` the most commonly used packages are `nlme` (Pinheiro *et al.*, 2015) which is for LMMs only, its extension to GLMMs with the function `glmmPQL` in the package `MASS` (Venables and Ripley, 2002) and the package `lme4` (Bates *et al.*, 2014). There are several differences between the packages: the estimation algorithm, the covariance structure for the random effects and whether they can include serial correlation. The software packages differ also with respect to the combination of multiple random effects they allow for, especially whether crossed random effects (*i. e.* each second level is observed within each first level random effect) and whether nested random effects (*i. e.* each second level random effect varies within each first level random effect) are possible.

Package:	`nlme`	`MASS`	`lme4`
Model:	LMM	GLMM	GLMM
Function name:	`lme`	`glmmPQL`	`glmer`
Algorithm:	IRLS	PQL-IRLS	P-IRLS
Marginal likelihood:	Laplace	Laplace	Gauss-Hermite or Laplace
Random effects:	nested only	nested only	nested and crossed
Covariance $\mathbf{D}(\boldsymbol{\delta})$:	generic	generic	diagonal or unstructured

Table 1.: Comparison of common `R` software packages for likelihood inference in GLMMs.

An overview for the comparison of these criteria is given in Table 1. Serial correlation models are only available for `nlme` and `glmmPQL`. The within-correlation can be generically defined by the user or a predefined correlation structure, such as an exponential correlation, can be used. In `lme4` the random effects covariance is assumed to be unstructured or diagonal *i. e.* uncorrelated, and has no possibility for serial correlation. Each package involves an IRLS algorithm with Fisher scoring based on similar numerical optimization routines. Nevertheless, the algorithms differ between packages: `nlme` updates the estimates for $\boldsymbol{\delta}$ to increase the accuracy, `glmmPQL` applies the PQL algorithm and `lme4` uses the P-IRLS algorithm. For the Gauss-Hermite approximation the number of quadrature points in `lme4` is hard-coded to a maximum of 25 since version 1.0-0. Only a Laplace approximation is available in `lme4` since version 1.0-0 for non-scalar random effects ($q > 1$), *e. g.* for a random intercept plus random slope model, or for two different random effect levels. Results with different software packages may differ substantially, although they implement the same estimation algorithm as discussed by Zhang *et al.* (2011).

3 Bayesian inference

Model parameters are treated as unknown random quantities in a Bayesian inference approach. This is in contrast to likelihood inference where model parameters are assumed to be true, fixed quantities. The basis for Bayesian inference is Bayes' theorem or Bayes' rule, which states how the conditional probability of an event can be reformulated as probability of the condition, given the event. The theorem is named after Reverend Thomas Bayes, whose work on probabilities of a binomial distribution was posthumously published in 1763. Bayes' rule for probabilities can be applied to $f(\cdot)$, a density function or probability mass function. It states that the so called posterior probability distribution of the unknown parameters $\boldsymbol{\theta}$ given the observed data $\mathbf{y}$ is

$$f(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{f(\mathbf{y} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta})}{f(\mathbf{y})}, \tag{8}$$

where $f(\mathbf{y}) = \int f(\mathbf{y} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta}$ is the marginal distribution function and in the case of discrete values for $\boldsymbol{\theta}$ is obtained by $f(\mathbf{y}) = \sum_{\boldsymbol{\theta}} f(\mathbf{y} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta})$. The marginal likelihood in the denominator in Equation (8) serves as normalizing constant which is independent of $\boldsymbol{\theta}$ such that one can write

$$f(\boldsymbol{\theta} \mid \mathbf{y}) \propto f(\mathbf{y} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta}). \tag{9}$$

The distribution of the parameters $f(\boldsymbol{\theta})$ is called prior distribution and $f(\mathbf{y} \mid \boldsymbol{\theta})$ is the sampling distribution, which in likelihood inference, under the assumption of $\boldsymbol{\theta}$ being fixed quantities, is just the likelihood $L(\boldsymbol{\theta}) = f(\mathbf{y} \mid \boldsymbol{\theta})$. Equation (9) states that the posterior distribution is proportional to the likelihood times the prior distribution. It also shows that if the prior distribution $f(\boldsymbol{\theta})$ is flat, *i. e.* uninformative, then the likelihood is just multiplied by a constant such that the posterior mode will coincide with the ML estimate. It also gets clear from Equation (9) that the influence of the prior relative to the likelihood decreases with increasing sample size. Adding observations will increase the product, or on the log-scale the sum, involved in the likelihood term $f(\mathbf{y} \mid \boldsymbol{\theta})$ and thus increase its relative weight, compared to the prior. The posterior distribution $f(\boldsymbol{\theta} \mid \mathbf{y})$ is the fundament for inference about $\boldsymbol{\theta}$. If one is interested in a particular model parameter θ_k then one examines the marginal posterior distribution

$$f(\theta_k \mid \mathbf{y}) = \int f(\mathbf{y} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta}_{-k} \tag{10}$$

where $\boldsymbol{\theta}_{-k}$ are all but the kth model parameter in $\boldsymbol{\theta}$. The marginal posterior can be used to obtain interval estimates, or point estimates for θ_k. This can *e. g.* be the posterior mean $E(\theta_k \mid \mathbf{y}) = \int \theta_k f(\theta_k \mid \mathbf{y}) d\theta_k$, the posterior mode $\text{Mode}(\theta_k \mid \mathbf{y}) = \arg\max_{\theta_k} f(\theta_k \mid \mathbf{y})$ or a credible interval $[t_l, t_u]$ with credible level $\text{CI}_\psi = \int_{t_l}^{t_u} f(\theta_k \mid \mathbf{y}) d\theta_k$. The lower and upper bound are equal to the quantiles $t_l = (1 - \psi)/2$ and $t_u = (1 + \psi)/2$ such that θ_k is within this interval with probability ψ. An introduction to Bayesian inference is given by Held and Sabanés Bové (2014, Chapter 6).

3.1 Posterior distributions for generalized linear mixed models

As mentioned in Section 1.2, the GLMM can be expressed as Bayesian hierarchical model. A Gaussian prior is assigned to the model parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{b})^\top$, with $\mathbf{b} \sim \text{N}(\mathbf{0}, \mathbf{D}(\boldsymbol{\delta}))$. The distribution of the random effects is the same as in likelihood inference in Section 2. A Gaussian prior for the fixed effects $\boldsymbol{\beta} \sim \text{N}(\mathbf{0}, \Sigma_{\boldsymbol{\beta}})$ is usually chosen such that the covariance matrix $\Sigma_{\boldsymbol{\beta}}$ is diagonal with very large and equal entries for each corresponding component of $\boldsymbol{\beta}$ such that essentially an uninformative prior results. The third layer in the hierarchical model

assigns a prior distribution to the hyperparameters of the latent Gaussian field $\boldsymbol{\theta}$, which are the parameters $\boldsymbol{\delta}$ that define the random effects covariance matrix and which will be included as additional factor in the computation of the posterior distribution. The model parameters $\boldsymbol{\theta}$ define a latent Gaussian field, for which the elements are conditionally independent, given the entity-specific stochastic dependence structure, such that a GMRF with a sparse precision matrix $\mathbf{Q}(\boldsymbol{\delta})$ results (Rue and Held, 2005, Chapter 4). The posterior distribution for the GLMM model parameters and hyperparameters, given the data $\mathbf{y}$ is

$$\begin{aligned} f(\boldsymbol{\theta}, \boldsymbol{\delta} \mid \mathbf{y}) &\propto f(\boldsymbol{\delta}) f(\boldsymbol{\theta} \mid \boldsymbol{\delta}) \prod_{i=1}^{I} f(\mathbf{y}_i \mid \boldsymbol{\theta}, \boldsymbol{\delta}) \\ &\propto f(\boldsymbol{\delta}) \mid \mathbf{Q}(\boldsymbol{\delta}) \mid^{\frac{1}{2}} \exp\left\{ -\frac{1}{2} \boldsymbol{\theta}^{\top} \mathbf{Q}(\boldsymbol{\delta}) \boldsymbol{\theta} + \sum_{i=1}^{I} \log f(\mathbf{y}_i \mid \boldsymbol{\theta}, \boldsymbol{\delta}) \right\} \end{aligned}$$

as discussed by Fong *et al.* (2010) who review the GLMM applications presented by Breslow and Clayton (1993) in the context of Bayesian inference for hierarchical latent Gaussian models. The marginal posterior distribution for a GLMM of the kth model parameter θ_k is then

$$\begin{aligned} f(\theta_k \mid \mathbf{y}) &= \int_{\boldsymbol{\delta}} \int_{\boldsymbol{\theta}_{-k}} f(\boldsymbol{\theta}, \boldsymbol{\delta} \mid \mathbf{y}) d_{\boldsymbol{\theta}_{-k}} d_{\boldsymbol{\delta}} \\ &= \int_{\boldsymbol{\delta}} f(\theta_k \mid \boldsymbol{\delta}, \mathbf{y}) f(\boldsymbol{\delta} \mid \mathbf{y}) d_{\boldsymbol{\delta}} \end{aligned} \tag{11}$$

and for the kth component of the hyperparameters the marginal posterior distribution is

$$f(\delta_k \mid \mathbf{y}) = \int_{\boldsymbol{\delta}_{-k}} f(\boldsymbol{\delta} \mid \mathbf{y}) d_{\boldsymbol{\delta}_{-k}} \tag{12}$$

where $\boldsymbol{\theta}_{-k}$ and $\boldsymbol{\delta}_{-k}$ are vectors with all components in the corresponding parameter vector except the kth one.

The possibilities to apply Bayesian inference used to be limited, as the integrals in Equation (10), or respectively in (11) and (12) and the summary statistics based on these marginal posterior distributions were only analytically solvable for selected problems. This was *e. g.* the case if likelihood and posterior were conjugate, *i. e.* posterior and prior belong to the same distribution family. From a Bayesian point of view Equation (5) in Section 2 for LMMs is in accordance with the desirable setting of having a conjugate prior distribution, namely a normal distribution for the likelihood and for the random effects, which results in an analytically solvable problem with a normal posterior distribution. The development of computers opened up the possibility of numerical integration. Coming along with the increase and availability of computing power, Bayesian inference experienced a boom during the nineties of the last century (Robert and Casella, 2011), using Markov chain Monte Carlo (MCMC) sampling. In the meantime also other strategies for evaluating integrals as in Equation (10) were established, such as the integrated nested Laplace approximations (INLA) (Rue *et al.*, 2009). Both methods, MCMC and INLA, are shortly introduced in the following two sections.

3.2 Markov chain Monte Carlo sampling

An introduction to numerical methods for Bayesian inference is provided by Held and Sabanés Bové (2014, Chapter 8). Robert and Casella (2011) give a short account about the roots and developments of modern MCMC techniques used in statistics. Numerical Monte Carlo

(MC) integration with computers was explored at the Los Alamos research center during the Second World War. An MC integration approximates *e.g.* the mean of the posterior $f(\theta \mid \mathbf{y})$, where θ is a scalar parameter. MC integration generates L independent random samples $\theta^{(1)}, \ldots, \theta^{(l)}, \ldots \theta^{(L)}$ from the posterior distribution and computes the mean as

$$E(\theta \mid \mathbf{y}) = \int \theta f(\theta \mid \mathbf{y}) d\theta \approx \frac{1}{L} \sum_{l=1}^{L} \theta^{(l)}$$

which converges to the true value $E(\theta \mid \mathbf{y})$ for $L \to \infty$. Similarly, one can construct estimates by using MC integration for other summary statistics based on the posterior distribution. Obtaining independent samples from the posterior distribution is difficult if there are many unknown model parameters $\boldsymbol{\theta}$. Sampling from the distribution of a high-dimensional vector $\boldsymbol{\theta}$ may result in large and persistent correlations between samples. A solution to this problem is to simulate a Markov chain $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(l)} \ldots, \boldsymbol{\theta}^{(L)}$, which generates samples $\boldsymbol{\theta}^{(l)}$ that depend only on the previous sample $\boldsymbol{\theta}^{(l-1)}$ and which converges to the posterior distribution $f(\boldsymbol{\theta} \mid \mathbf{y})$. Given that the Markov chain converged to the posterior distribution one can again apply MC integration to obtain the summary statistics of interest. The combination of the two eponymous procedures defines MCMC sampling. The Metropolis Hastings algorithm (Metropolis *et al.*, 1953; Hastings, 1970) describes how a Markov chain can be generated such that it converges to the posterior distribution. Starting with some values for model parameter θ_k a proposal $\theta_k^\star$ is defined by drawing randomly from the proposal density $f^\star\left(\theta_k^\star \mid \theta_k, \boldsymbol{\theta}_{-k}\right)$ which depends only on the current value $\boldsymbol{\theta}^{(l)}$ of the simulated Markov chain. The parameter is updated to $\theta_k^{(l+1)} = \theta_k^\star$ with acceptance probability α equal to

$$\alpha = \min \left\{ 1, \frac{f\left(\theta_k^\star \mid \boldsymbol{\theta}_{-k}, \mathbf{y}\right)}{f\left(\theta_k^{(l)} \mid \boldsymbol{\theta}_{-k}, \mathbf{y}\right)} \frac{f^\star\left(\theta_k^{(l)} \mid \theta_k^\star, \boldsymbol{\theta}_{-k}\right)}{f^\star\left(\theta_k^\star \mid \theta_k^{(l)}, \boldsymbol{\theta}_{-k}\right)} \right\}$$

and otherwise $\theta_k^{(l+1)} = \theta_k^{(l)}$. Each model parameter θ_k in $\boldsymbol{\theta}$ can be updated, conditional on all other current model parameters $\boldsymbol{\theta}_{-k}$ with some proposal density $f^\star(\cdot)$ and the MH algorithm will converge to the posterior distribution if L is large enough.

The Gibbs sampler, introduced by Geman and Geman (1984) and later discussed by Gelfand and Smith (1990), modifies the MH algorithm by setting $f^\star\left(\theta_k^\star \mid \theta_k^{(l)}, \boldsymbol{\theta}_{-k}^{(l)}\right) = f\left(\theta_k^\star \mid \boldsymbol{\theta}_{-k}, \mathbf{y}\right)$ *i.e.* the proposal density is equal to the target posterior density. The Gibbs algorithm thus samples component-wise from the full conditionals of every model parameter θ_k and has an acceptance probability equal to one. Instead of component-wise updating every θ_k one can use block-updating schemes (Rue and Held, 2005, section 4.1.2), which is preferable if model parameters are highly correlated, which is discussed by Gamerman (1997) in the context of GLMMs. The hypothesis that the generated Markov chain converged to a stationary posterior distribution must be examined for every model parameter by *e.g.* visual inspection of the trace plots, examination of the autocorrelation function of the samples or checking different convergence diagnostics (Cowles and Carlin, 1996).

3.3 Integrated nested Laplace approximations

An alternative to MCMC sampling was proposed by Rue *et al.* (2009) and is called integrated nested Laplace approximations (INLA) which is an approximate Bayesian approach. An introduction to INLA is provided by Blangiardo and Cameletti (2015, in Chapter 4). INLA

approximates the marginal posterior distribution $f(\theta_k \mid \mathbf{y}, \boldsymbol{\delta})$ for a Bayesian hierarchical model with a latent Gaussian field that follows a GMRF and which has relatively few, say less than six, hyperparameters according to Rue *et al.* (2009).

The first task is to approximate the joint distribution of all hyperparameters $f(\boldsymbol{\delta} \mid \mathbf{y})$, which appear in Equation (11) and from which the marginals in Equation (12) can be derived. The distribution of the hyperparameters is approximated by

$$f(\boldsymbol{\delta} \mid \mathbf{y}) = \frac{f(\boldsymbol{\theta}, \boldsymbol{\delta} \mid \mathbf{y})}{f(\boldsymbol{\theta} \mid \boldsymbol{\delta}, \mathbf{y})} \propto \frac{f(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\delta}) f(\boldsymbol{\theta} \mid \boldsymbol{\delta}) f(\boldsymbol{\delta})}{f(\boldsymbol{\theta} \mid \boldsymbol{\delta}, \mathbf{y})} \approx \left. \frac{f(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\delta}) f(\boldsymbol{\theta} \mid \boldsymbol{\delta}) f(\boldsymbol{\delta})}{\tilde{f}_G(\boldsymbol{\theta} \mid \boldsymbol{\delta}, \mathbf{y})} \right|_{\boldsymbol{\theta} = \boldsymbol{\theta}^\star(\boldsymbol{\delta})} = \tilde{f}(\boldsymbol{\delta} \mid \mathbf{y}) \tag{13}$$

where $\tilde{f}_G(\boldsymbol{\theta} \mid \boldsymbol{\delta}, \mathbf{y})$ is a Gaussian approximation to the full conditional of $\boldsymbol{\theta}$ evaluated at the mode $\boldsymbol{\theta}^\star(\boldsymbol{\delta})$ for a given $\boldsymbol{\delta}$. The approximation $\tilde{f}(\boldsymbol{\delta} \mid \mathbf{y})$ corresponds to the Laplace approximation to marginal posteriors discussed by Tierney and Kadane (1986). The proportionality in (13) is with respect to the normalizing constant $f(\mathbf{y})$ in the posterior distribution $f(\boldsymbol{\theta}, \boldsymbol{\delta} \mid \mathbf{y})$.

The approximation to the second term in Equation (11) can be done with three different methods, with different levels of accuracy. The first, simplest and least accurate approach is to use $\tilde{f}_G(\boldsymbol{\theta} \mid \boldsymbol{\delta}, \mathbf{y})$ and derive a normal distribution to approximate each marginal distribution $\tilde{f}_G(\theta_k \mid \boldsymbol{\delta}, \mathbf{y})$ using the mean of the Gaussian approximation and the marginal variance. As there may be errors due to shifts in location or due to skewness (Rue and Martino, 2007), a second Laplace approximation to the marginal $f(\theta_k \mid \boldsymbol{\delta}, \mathbf{y})$ by the approach of Tierney and Kadane (1986) results in a higher accuracy.

This full Laplace approximation is obtained by

$$f(\theta_k \mid \boldsymbol{\delta}, \mathbf{y}) = \frac{f(\theta_k, \boldsymbol{\theta}_{-k} \mid \boldsymbol{\delta}, \mathbf{y})}{f(\boldsymbol{\theta}_{-k} \mid \theta_k, \boldsymbol{\delta}, \mathbf{y})} = \frac{f(\boldsymbol{\theta}, \boldsymbol{\delta} \mid \mathbf{y})}{f(\boldsymbol{\delta} \mid \mathbf{y})} \frac{1}{f(\boldsymbol{\theta}_{-k} \mid \theta_k, \boldsymbol{\delta}, \mathbf{y})}$$

$$\propto \frac{f(\boldsymbol{\theta}, \boldsymbol{\delta}, \mathbf{y})}{f(\boldsymbol{\theta}_{-k} \mid \theta_k, \boldsymbol{\delta}, \mathbf{y})} \approx \left. \frac{f(\boldsymbol{\theta}, \boldsymbol{\delta}, \mathbf{y})}{\tilde{f}_G(\boldsymbol{\theta}_{-k} \mid \theta_k, \boldsymbol{\delta}, \mathbf{y})} \right|_{\boldsymbol{\theta}_{-k} = \boldsymbol{\theta}^\star_{-k}(\theta_k, \boldsymbol{\delta})} = \tilde{f}(\theta_k \mid \boldsymbol{\delta}, \mathbf{y}) \tag{14}$$

where $\tilde{f}_G(\boldsymbol{\theta}_{-k} \mid \theta_k, \boldsymbol{\delta}, \mathbf{y})$ is a Gaussian approximation to the full conditional $f(\boldsymbol{\theta}_{-k} \mid \theta_k, \boldsymbol{\delta}, \mathbf{y})$ evaluated at the mode $\boldsymbol{\theta}^\star_{-k}$ of the full conditional for a given $\boldsymbol{\theta}_{-k}$ and given $\boldsymbol{\delta}$. The approximation $\tilde{f}_G(\boldsymbol{\theta}_{-k} \mid \theta_k, \boldsymbol{\delta}, \mathbf{y})$ requires a high computational effort as it needs to be evaluated for each entity in the GMRF and for each $\boldsymbol{\delta}$. Thus Rue *et al.* (2009) suggest two simplifications. First, they propose to use conditional densities derived from the already computed approximation $\tilde{f}_G(\boldsymbol{\theta} \mid \boldsymbol{\delta}, \mathbf{y})$ from Equation (13) to approximate the mode $\boldsymbol{\theta}^\star_{-k}(\theta_k, \boldsymbol{\delta})$. Secondly, they restrict the influence of $\boldsymbol{\theta}_{-k}$ on the approximation for θ_k, as the dependence between two entities in the GMRF is assumed to decay with increasing distance. With these two simplifications the Laplace approximation $\tilde{f}(\theta_k \mid \boldsymbol{\delta}, \mathbf{y})$ corresponds to the Gaussian approximation multiplied by a term which is equivalent to a cubic spline for each entity-specific parameter $\boldsymbol{\theta}_k$.

The accuracy and computational costs of the third approximation, called simplified Laplace approximation, is between the simple Gaussian and the more precise full Laplace approximation. The simplified Laplace approximation applies a Taylor series expansion up to the third order to the nominator $f(\boldsymbol{\theta}, \boldsymbol{\delta}, \mathbf{y})$ and to the denominator $\tilde{f}_G(\boldsymbol{\theta}_{-k} \mid \theta_k, \boldsymbol{\delta}, \mathbf{y})$ in Equation (14) around the mean for entity k which is derived from $\tilde{f}_G(\boldsymbol{\theta} \mid \boldsymbol{\delta}, \mathbf{y})$. This adds a correction term to the Gaussian approximation for location and skewness and reduces the computational costs to approximate $\tilde{f}(\theta_k \mid \boldsymbol{\delta}, \mathbf{y})$, compared to the full Laplace approximation which involves the cubic spline term for every entity-specific model parameter $\boldsymbol{\theta}_k$.

For INLA both terms, $\tilde{f}(\boldsymbol{\delta} \mid \mathbf{y})$ and $\tilde{f}(\theta_k \mid \boldsymbol{\delta}, \mathbf{y})$, are used to numerically integrate over different integration points $\boldsymbol{\delta}_u$ and different weights Δ_u

$$\tilde{f}(\theta_k \mid \mathbf{y}) \approx \sum_u \tilde{f}(\theta_k \mid \boldsymbol{\delta}_u, \mathbf{y}) \tilde{f}(\boldsymbol{\delta}_u \mid \mathbf{y}) \Delta_u \tag{15}$$

to get the approximated marginal posterior distribution $\tilde{f}(\theta_k \mid \mathbf{y})$. The choice of the points $\boldsymbol{\delta}_u$ and the weights Δ_u is discussed in more detail in Section 3.5.

INLA was demonstrated to deliver accurate approximations of the marginal posterior distributions coming with reduced computational costs compared to MCMC. See Rue *et al.* (2009) for examples or Schrödle *et al.* (2011) for applications to spatio-temporal models. Lindgren *et al.* (2011) illustrate the numerical advantages of sparse matrices implied by GMRFs for large geostatistical models which are solved fast and accurately by INLA. For the applicability of INLA to GLMMs see Rue *et al.* (Section 5.2 2009) and Fong *et al.* (2010).

3.4 Choice of prior distribution

Selecting a prior distribution for $f(\boldsymbol{\theta})$ and disclose its influence on the posterior distribution is one of the most disputed elements in Bayesian inference. Gaining new knowledge about $\boldsymbol{\theta}$ based on the inclusion of prior beliefs may come with the flavour of being subjective, *i. e.* biased and was criticised beyond the field of statistics (Popper, 1959). On the other hand, Bayes' theorem offers a rationale on how historical data, which was observed and perhaps should not be ignored, could be taken into account and how to evaluate it in the context of new evidence. Historical data from similar previous studies may be available, which is rather common for clinical trials and could serve as prior information, also by introducing a prior weight on the historical data directly, like suggested by Ibrahim and Chen (2000).

Depending on the choice of the prior distribution, one may establish links between Bayesian inference and a ML approach. For instance, a Bayesian interpretation of the REML in LMMs is discussed by Harville (1974). One could choose a non-informative, flat prior which is proportional to a constant on the fixed effects $f(\boldsymbol{\beta}) \propto c$ and as well for the hyperpameters $f(\boldsymbol{\delta}) \propto c$. The mode of the joint posterior distribution with respect to $\boldsymbol{\delta}$ is in this case equivalent to the ML estimate of the hyperparameters. In contrast, the mode of the the marginal posterior distribution with respect to $\boldsymbol{\delta}$ is equivalent to the REML estimate of $\boldsymbol{\delta}$, which is also mentioned in Section 2.1.

Instead of assigning a prior distribution on the hyperparameters one could assume unknown and fixed values for $\boldsymbol{\delta}$. An estimate for $\boldsymbol{\delta}$ could be obtained by maximizing the marginal likelihood, which results in the REML estimate for $\boldsymbol{\delta}$. Using this REML estimate of the hyperpameters to analyse the posterior mean of $\boldsymbol{\beta}$ and $\mathbf{b}$ results in the same EBLUP estimates for $\boldsymbol{\beta}$ and $\mathbf{b}$ as in Section 2.1. This approach, without assigning a prior distribution to $\boldsymbol{\delta}$, is called empirical Bayes. Also for GLMMs the posterior modes, based on Bayesian inference with empirical Bayes, correspond to the ML estimates.

A fully Bayesian approach, in contrast to empirical Bayes, assigns a prior distribution to all parameters $\boldsymbol{\theta}$ and additionally to the hyperparameters $\boldsymbol{\delta}$. The choice of a non-informative, improper prior for $\boldsymbol{\delta}$, which does not integrate to unity, does not guarantee in general that the posterior distribution will be proper. This holds also especially for Jeffreys' prior, which is invariant to a reparametrisation. Usually one resorts to choose a weakly informative prior instead. In the case of one single random effect, say a random intercept, only a single hyperparameter $\delta = \sigma^2_{\text{RI}}$ results, which is the random intercept variance. An inverse gamma prior $\sigma^2_{\text{RI}} \sim \text{IG}(\alpha_1, \alpha_2)$ is conjugate (Held and Sabanés Bové, 2014, Chapter 6.3.3) to the normal dis-

tribution of $\mathbf{b}$, meaning that $f(\sigma^2_{\text{RI}} \mid \mathbf{b})$ again belongs to an inverse gamma distribution. One could of course choose the parameters α_1 and α_2, such that the prior is only weakly informative. The extension of this conjugate prior to the case where $q > 1$, *e. g.* a random intercept plus slope model, with a $q \times q$ random effects covariance matrix, leads to an inverse Wishart distribution (Held and Sabanés Bové, 2014) for $\mathbf{D}(\boldsymbol{\delta})$. Fong *et al.* (2010) motivate the choice of an informative prior in GLMMs for such an inverse gamma or inverse Wishart distribution.

However, the inverse gamma prior on random effect variances like σ^2_{RI} was found by Roos and Held (2011) to result in a large sensitivity for the parameter estimates. As an alternative, they propose to use a half-normal prior distribution on the standard deviation, which is also suggested by Gelman (2006). Gelman *et al.* (2008) discusses how to assess a weakly informative default prior in the context of hierarchical models.

On the other hand, there may occur situations, such as sparse data, for which an explicit informative prior on $\boldsymbol{\beta}$ may be favourable (Greenland, 2006). The choice of an informative prior for $\boldsymbol{\delta}$ affects the amount of shrinkage for the estimates of $\mathbf{b}_i$. In Section 2 the estimates were asserted to be shrunk towards the population averaged mean response profile. As the posterior is obtained by multiplying likelihood and prior distribution, the location and the amount of shrinkage for $\mathbf{b}_i$ can directly be influenced by choosing the moments for the prior distribution which is assigned to $\boldsymbol{\delta}$. The problem of (quasi) complete separation, mentioned in Section 2, may be put into the context of a sparse data problem (Firth, 1993). In the case of a binary covariate in a binomial GLM, where the ML estimates are not defined, complete separation arises if the off diagonal entries of the corresponding 2×2 contingency table are zero. According to Firth (1993) this can be addressed by a penalized likelihood, for which the penalization term depends on the inverse Fisher information and is related to Jeffreys' invariant prior. For a logistic regression with a completely separating binary covariate this approach corresponds to adding $1/2$ to each cell of the 2×2 table. A penalization term is in this situation related to a Bayesian approach which assigns an informative prior distribution, based on which the implied shrinkage may help to solve the problem of a non-existent ML estimate in a consistent way.

Similarly, according to Greenland (2006, 2007a,b, 2009) the use of proper, informative priors may help *e. g.* in epidemiological studies with few data, to avoid possibly unrealistic assumptions of a likelihood inference approach. For the fixed effects in a Bayesian hierarchical regression, the normal prior $\boldsymbol{\beta} \sim \text{N}(\mathbf{0}, \Sigma_{\boldsymbol{\beta}})$, is usually chosen such that $\Sigma_{\boldsymbol{\beta}}$ is diagonal with large values for the corresponding variances. An informative prior would motivate smaller variances for some components of $\boldsymbol{\beta}$ and possibly also deviate form the mean zero location parameter. In Paper IV such informative priors are proposed for GLMs as well as for GLMMs, which is in line with the motivation by Greenland (2006). Furthermore, in Paper IV also the diagonal structure for $\Sigma_{\boldsymbol{\beta}}$ is relaxed and a prior weight, based on the observed correlations in the data, is used for $\boldsymbol{\beta}$. These adaptive prior weights are perfectly treatable with the INLA approach and were implemented by using the `r-inla` software. Motivating an informative prior or using default or reference priors will not circumvent the indispensable questions about the impact of the prior and the sensitivity of the estimates with respect to alternative prior specifications. Investigating prior sensitivity becomes attractive with computationally less intensive methods such as INLA (Roos and Held, 2011; Roos *et al.*, 2015), compared to the prohibitive computational costs inflicted by MCMC. Any disagreement between the observed data and the chosen prior, which implies a relatively strong influence of the prior on the posterior estimates, can be disclosed by *e. g.* Box's-p value (Box, 1980), which is also discussed in Paper IV.

3.5 Software

There also exist several software packages for Bayesian inference in GLMMs. In the following, a short overview is given for generic MCMC samplers and the `r-inla` package, which implements the INLA approach discussed in Section 3.3.

MCMC

After the potential of MCMC sampling for Bayesian inference was recognized it did not last long until efforts for a common computer language which implements generic Gibbs samplers and which allows for a broad set of different applications, were initiated (Gilks *et al.*, 1994). This resulted in the BUGS (Bayesian Inference Using Gibbs Sampling) project (Lunn *et al.*, 2009), which defined a program language for generic MCMC samplers like Win-BUGS and Open-BUGS (Lunn *et al.*, 2000), JAGS (Plummer, 2003). They all have a common syntax to set up Bayesian hierarchical models. These generic MCMC samplers can be used for hierarchical models with several layers of parameter levels and are not restricted to latent Gaussian fields. There are several interfaces, like the `R` package `R2jags` or `coda` among others, which provide access to the flexible `R` environment and a collection of specific functions to analyse MCMC sampling results. Another common software package in `R` is `MCMCglmm` (Hadfield, 2010) which is a MCMC sampler for multivariate GLMMs, uses a similar syntax as the package `nlme` and allows for correlated random effects. Yet another software package, called STAN Gelman *et al.* (2014); Stan Development Team (2014), uses a distinct modelling language and different methods for MCMC sampling. Most of these MCMC samplers implement a Gibbs sampler, or a general Metropolis Hastings algorithm. Nevertheless, there may be crucial differences in the implemented algorithms, for example if it comes to block updating, where *e. g.* JAGS uses the algorithm proposed by Holmes and Held (2006).

INLA

The `R` package `r-inla` (Rue *et al.*, 2014), which is available on `http://www.r-inla.org`, is essentially an interface to the standalone package `INLA` which in turn calls the `GMRFLib` library (Rue and Held, 2005) which is written in C and Fortran. The `r-inla` package defines models with a similar syntax like the established `glm` regression model framework in `R` and offers the possibility to process results by the flexible facilities of the `R` environment.

Gaussian approximations to the marginal posterior (Tierney and Kadane, 1986) are obtained in `r-inla` by a Fisher scoring algorithm based on numerical optimization routines. The approximation to the joint posterior of the hyperparameters $f(\boldsymbol{\delta} \mid \boldsymbol{\theta}, \mathbf{y})$ involves three steps: first the mode is searched by a quasi-Newton method involving differences between gradients. The second step evaluates the curvature at the mode to get the Fisher information matrix and for which an Eigen decomposition is computed. Based on the standardized, orthogonal components the approximated posterior $\tilde{f}(\boldsymbol{\delta} \mid \mathbf{y})$ is explored. The points $\boldsymbol{\delta}_u$ in Equation (15) at which $\tilde{f}(\boldsymbol{\delta} \mid \mathbf{y})$ is explored can be determined by two different strategies: the first places a grid of 'interesting' points around the mode in each direction of the standardized variables with a certain step-length, as long as the difference in the log-densities does not exceed a stopping criterion. The second integration strategy, the central composite design (CCD) (Rue *et al.*, 2009, Section 6.5), explores the posterior density with less points, thus is less accurate and requires less computational effort. This second integration strategy exploits the curvature in the Fisher information matrix to determine a sphere around the mode, at which the evaluation points are chosen. The corresponding weights Δ_u in Equation (15) are determined depending

on the selected integration strategy. The grid integration uses equal weights and determines the new points depending on $\boldsymbol{\delta}_{\boldsymbol{u}}$, which are the points that were already explored. The CCD integration adapts Δ_u depending on the radius of the computed sphere.

In the `r-inla` package there are control parameters to set the step-length for the gradient calculations in finding the mode, the step-length for finding the points on the standardized scale at which $\tilde{f}(\boldsymbol{\delta} \mid \mathbf{y})$ is explored and the difference in the log-densities if the grid integration strategy is chosen. Each marginal posterior $\tilde{f}(\delta_k \mid \mathbf{y})$ is obtained by interpolating between the points at which the joint posterior $\tilde{f}(\boldsymbol{\delta} \mid \mathbf{y})$ was already explored.

The `r-inla` package provides a set of implemented likelihood functions for $f(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\delta})$ which fit into the GLM framework. A collection of GMRFs, such as a random walk, an autoregressive or an independent model but also models for spatial dependencies or a generic GMRF for which a user-specified covariance matrix can be provided, are implemented in `r-inla`. A combination of these implemented latent fields, as denoted by the additive functional terms in Equation (4), is possible. Each functional term, or GMRF, can have weights and the point in the different latent fields may be correlated. A simple example for such a model is a random intercept and random slope model for longitudinal data: both random effects define a functional component for each entity and both random effects are correlated within each entity where the random slope use the time covariate as weight. For the hyperparameters $f(\boldsymbol{\delta})$ there is also a collection of ready to use priors implemented, but a user-specified prior of suitable form for each hyperparameter can be passed as argument in the form of a tabulated vector. It is possible to evaluate also linear combinations of the marginal posteriors, although they must be defined in advance. Furthermore it is possible to use a model with multivariate response in `r-inla`, as for example discussed by Paul *et al.* (2010), which just requires that the response matrix and the covariates are organized in an adequate form. The software allows for the use of different likelihoods in the case of a multivariate response model, as discussed by Martino *et al.* (2011) who use two likelihoods for a joint model for longitudinal and survival data, or illustrated by Muff *et al.* (2015) in the context of a measurement error model. Another feature of `r-inla` is that one can relax the assumption that every entity-specific response $\mathbf{y}_i$ is related to one single entity in the latent field. A correlation in the latent field across different entities is possible, as discussed Riebler *et al.* (2012) who correlate the multivariate response of different countries in an age-period-cohort model. An overview about these special features in `r-inla` is given by Martins *et al.* (2013).

INLA is an approximate Bayesian inference approach and comes with a certain analytically not assessable error (see the discussion in Rue *et al.*, 2009). MCMC sampling requires the user to assess if the stationary distribution was reached for the obtained samples. Additionally, MCMC sampling also comes with a sampling error, as the convergence property only holds asymptotically. The key advantage of INLA compared to MCMC sampling, which for a long enough Markov chain converges to the true distribution, is its comparably low computational cost, which especially plays out in complex models. The broad model class of latent Gaussian fields together with the `r-inla` software package makes it possible to set up various kinds of different models by using distributions from the exponential family for the outcome, setting up own priors for the hyperparameters and offering a range of features, such as the correlation across entities in the latent Gaussian field, or by allowing for multiple likelihoods, which makes `r-inla` attractive for a diverse set of applications.

Thesis Summary

This thesis consists of four papers. The four projects evolved at different paces during the last three years. They all have in common that they deal with GLMMs and in all papers the discussed statistical methods are illustrated with epidemiological applications. Paper I uses likelihood inference for a longitudinal data analysis with an LMM. The second, Paper II, discusses the inaccuracy encountered with INLA in the special case of cluster-specific quasi-complete separation in a binary response GLMM. Paper III discusses the application of INLA to a special form of GLMM, namely in the context of a network meta-analysis. Paper IV introduces adaptive prior weights for the fixed effects in a GLMM, which are possibly correlated. The content of each paper is briefly summarised below.

Paper I

CD8 counts and CD4/CD8 ratio independently predict CD4 response in drug naive and in patients on cART Rafael Sauter, Ruizhu Huang, Bruno Ledergerber, Manuel Battegay, Enos Bernasconi, Matthias Cavassini, Hansjakob Furrer, Matthias Hoffmann, Mathieu Rougemont, Huldrych F. Günthard, Leonhard Held & the Swiss HIV cohort study.

This paper analyses if the CD8 lymphocyte subtype is a predictor for the HIV disease progression which is measured by CD4 lymphocyte cell counts. The paper analyses this hypothesis based on longitudinal data from the Swiss HIV cohort study with a LMM and uses likelihood inference methods. Other studies investigated extensively the relationship between CD4 lymphocyte cell counts and the viral load, measured as number of RNA copies per blood volume. However, there seems to be a lack of knowledge about the relationship between CD4 and CD8 lymphocyte subtypes among HIV-1 infected patients.

The research question was proposed by Prof. Huldrych Günthardt from the University Hospital, Zurich and was financially supported by the SHCS research council. A large part of the data preparation and initial inference was part of the master thesis by Rhuizu Huang and was supervised by Leonhard Held and myself. The final analyses and the writing of the manuscript were done by myself. Leonhard Held, Bruno Ledergerber, Huldrych Günthardt and Ruizhu Huang reviewed the manuscript at several stages and contributed to improve it. The remaining co-authors read and contributed comments to the final version of the paper. The main contribution of this paper is the new insight about the dependencies between CD4 and CD8 lymphocyte subtypes among HIV-1 infected patients based on the SHCS data.

Paper II

Quasi-complete Separation in Random Effects of Binary Response Mixed Models: Integrated Nested Laplace Approximations vs. MCMC by Rafael Sauter, Leonhard Held.

This paper compares the results for GLMMs with a binary response outcome obtained by INLA to the results based on MCMC sampling and ML estimation. Initially the intention was to examine the possibility to use a correlated random effect structure, as proposed by Riebler *et al.* (2012), for longitudinal data. Based on the toenail data example, which repeatedly served as example for methodological discussions (Lesaffre and Spiessens, 2001; Verbeke and Molenberghs, 2013), I found that INLA can not reproduce the results obtained by MCMC nor by ML estimation.

The relatively large inaccuracy of INLA in the case of binary responses was already mentioned

by Rue *et al.* (2009). Fong *et al.* (2010) documented in a simulation study for binary response GLMMs the degree of the error for INLA, which was found to be acceptable by Grilli *et al.* (2014) if an informative prior is chosen. The particular problem in the toenail dataset goes back to a patient-specific quasi complete separation for the random effect parameters, as there is a large proportion of patients who always stay in the same response state. This paper illustrates, based on the toenail example and a simulation study, that INLA may be rather inaccurate depending on the degree of quasi-complete separation in the data, exceeding the degree of the error reported earlier (Fong *et al.*, 2010; Grilli *et al.*, 2014). I did the computations and wrote the manuscript, which was read and amended by Leonhard Held at several stages.

Paper III

Network meta-analysis with integrated nested Laplace approximations by Rafael Sauter, Leonhard Held.
This paper illustrates that it is possible to carry out network-meta analyses (NMA) as discussed by Lumley (2002) and by Lu and Ades (2006) with INLA, taking into account possible heterogeneity between trials and inconsistencies in the network through random effects. Also the node splitting approach suggested by Dias *et al.* (2010) is carried out with INLA. The idea to carry out a NMA with INLA goes back to Prof. Martin Schumacher of the University of Freiburg. I implemented several examples for different NMAs and the node-splitting approach in INLA and discussed the special features one needs to address, such that the NMAs fit the requirements by INLA. The manuscript was written by myself and Leonhard Held, who also contributed to the work by regularly discussing intermediate results and providing various suggestions for improvements. The main contribution of this paper is that it illustrates that NMA and especially node-splitting is possible with INLA, which allows to apply this method to large networks which were found to be computationally too involved to be addressed by MCMC sampling (Veroniki *et al.*, 2013).

Paper IV

Adaptive prior weighting in generalized linear models by Leonhard Held, Rafael Sauter.
This paper investigates possible prior-data conflicts in a regression model based on Box p-value and suggests to use adaptive prior weights, by including the information of the observed data. The prior weights can be introduced for the joint prior distribution of the regression coefficients, for independent coefficients or for specified blocks. The method is implemented in INLA and it is illustrated how to use adaptive prior weights for GLM's and how to extend this to GLMMs.

This paper addresses the use of informative priors in a situation of sparse data. Informative priors avoid unrealistic estimates as discussed by Greenland (2006, 2007a,b, 2009). The informative priors are constructed by using adaptive weights based on the collected observations and possible prior-data conflicts are discussed.

Leonhard Held had the idea to examine adaptive prior weighting in this context and he wrote the manuscript to which I contributed comments. I implemented the applications in INLA with the inspiration of code for a related analysis Held *et al.* (2012), written and kindly provided by Daniel Sabanés Bové. I wrote the supplementary material to the paper, which was reviewed by Leonhard Held.

References

Airy, G. B. (1861). *On the Algebraical and Numerical Theory of Errors of Observations and the Ccombination of Observations*, London, Macmillan.

Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models, *Biometrika* **71**(1): 1–10. Available from: http://biomet.oxfordjournals.org/content/71/1/1.abstract.

Ashford, J. R. and Sowden, R. R. (1970). Multi-variate probit analysis, *Biometrics* **26**(3): pp. 535–546. Available from: http://www.jstor.org/stable/2529107.

Bates, D. M. and DebRoy, S. (2004). Linear mixed models and penalized least squares, *Journal of Multivariate Analysis* **91**(1): 1 – 17. Special Issue on Semiparametric and Nonparametric Mixed Models. Available from: http://www.sciencedirect.com/science/article/pii/S0047259X04000867.

Bates, D., Maechler, M., Bolker, B. and Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-7. Available from: http://CRAN.R-project.org/package=lme4.

Blangiardo, M. and Cameletti, M. (2015). *Spatial and Spatio-temporal Bayesian Models with R-INLA*, John Wiley & Sons, Ltd.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness, *Journal of the Royal Statistical Society. Series A (General)* **143**(4): 383–430. Available from: http://www.jstor.org/stable/2982063.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association* **88**(421): 9–25. Available from: http://www.jstor.org/stable/2290687.

Breslow, N. E. and Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion, *Biometrika* **82**(1): pp. 81–91. Available from: http://www.jstor.org/stable/2337629.

Cowles, M. K. and Carlin, B. P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review, *Journal of the American Statistical Association* **91**(434): pp. 883–904. Available from: http://www.jstor.org/stable/2291683.

Cox, D. R. (1972). The analysis of multivariate binary data, *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **21**(2): pp. 113–120. Available from: http://www.jstor.org/stable/2346482.

Dias, S., Welton, N. J., Caldwell, D. M. and Ades, A. E. (2010). Checking consistency in mixed treatment comparison meta-analysis, *Statistics in Medicine* **29**(7-8): 932–944. Available from: http://dx.doi.org/10.1002/sim.3767.

Diggle, P. (2002). *Analysis of Longitudinal Data*, Oxford Statistical Science Series, Oxford University Press.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2 edn, Springer-Verlag New York.

Fahrmeir, L., Kneib, T. and Lang, S. (2004). Penalized structured additive regression for space-time data: A Bayesian perspective, *Statistica Sinica* **14**(3): 715–745.

Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. (2013). *Regression - Models, Methods and Applications*, Springer.

Firth, D. (1993). Bias reduction of maximum likelihood estimates, *Biometrika* **80**(1): 27–38. Available from: `http://www.jstor.org/stable/2336755`.

Fisher, R. A. (1918). The correlation between relatives on the supposition of mendelian inheritance., *Transactions of the Royal Society of Edinburgh* **52**: 399–433. Available from: `http://journals.cambridge.org/article_S0080456800012163`.

Fisher, R. A. (1925). *Statistical methods for research workers*, Cosmo Publications.

Fitzmaurice, G., Laird, N. and Ware, J. (2004). *Applied Longitudinal Analysis*, Wiley Series in Probability and Statistics, Wiley-Interscience.

Fitzmaurice, G., M., D., Verbeke, G. and Molenberghs, G. (2008). *Longitudinal Data Analysis*, Chapman & Hall.

Fong, Y., Rue, H. and Wakefield, J. (2010). Bayesian inference for generalized linear mixed models, *Biostatistics* **11**(3): 397–412.

Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear mixed models, *Statistics and Computing* **7**(1): 57–68.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association* **85**(410): 398–409.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper), *Bayesian Analysis* **1**(3): 515–534. Available from: `http://dx.doi.org/10.1214/06-BA117A`.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, Donald, B. (2014). *Bayesian Data Analysis*, 3 edn, Chapman & Hall.

Gelman, A., Jakulin, A., Pittau, M. G. and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models, *Annals of Applied Statistics* **2**(4): 1360–1383.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEE Transactions on Pattern Analysis and Machine Intelligence* **6**: 721–741.

Gilks, W. R., Thomas, A. and Spiegelhalter, D. J. (1994). A language and program for complex Bayesian modelling, *Journal of the Royal Statistical Society. Series D (The Statistician)* **43**(1): pp. 169–177. Available from: `http://www.jstor.org/stable/2348941`.

Gilmour, A. R., Anderson, R. D. and Rae, A. L. (1985). The analysis of binomial data by a generalized linear mixed model, *Biometrika* **72**(3): pp. 593–599. Available from: `http://www.jstor.org/stable/2336731`.

Greenland, S. (2006). Bayesian perspectives for epidemiological research: I. Foundations and basic methods, *International Journal of Epidemiology* **35**: 765–775.

Greenland, S. (2007a). Bayesian perspectives for epidemiological research. II. Regression analysis, *International Journal of Epidemiology* **36**(1): 195–202. Available from: http://ije.oxfordjournals.org/content/36/1/195.abstract.

Greenland, S. (2007b). Prior data for non-normal priors, *Statistics in Medicine* **26**(19): 3578–3590. Available from: http://dx.doi.org/10.1002/sim.2788.

Greenland, S. (2009). Bayesian perspectives for epidemiologic research: III. Bias analysis via missing-data methods, *International Journal of Epidemiology* **38**(6): 1662–1673. Available from: http://ije.oxfordjournals.org/content/38/6/1662.abstract.

Grilli, L., Metelli, S. and Rampichini, C. (2014). Bayesian estimation with integrated nested Laplace approximation for binary logit mixed models, *Journal of Statistical Computation and Simulation* pp. 1–9. Available from: http://dx.doi.org/10.1080/00949655.2014.935377.

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package, *Journal of Statistical Software* **33**(2): 1–22. Available from: http://www.jstatsoft.org/v33/i02/.

Harville, D. (1974). Bayesian inference for variance components using only error contrasts, *Biometrika* **61**(2): 383–385. Available from: http://biomet.oxfordjournals.org/content/61/2/383.abstract.

Harville, D. (1976). Extension of the Gauss-Markov theorem to include the estimation of random effects, *The Annals of Statistics* **4**(2): pp. 384–395. Available from: http://www.jstor.org/stable/2958209.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems, *Journal of the American Statistical Association* **72**(358): pp. 320–338. Available from: http://www.jstor.org/stable/2286796.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, Chapman & Hall.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* **57**(1): pp. 97–109. Available from: http://www.jstor.org/stable/2334940.

Held, L. and Sabanés Bové, D. (2014). *Applied Statistical Inference; Likelihood and Bayes*, Springer.

Held, U., Sabanés Bové, D., Steurer, J. and Held, L. (2012). Validating and updating a risk model for pneumonia - a case study, *BMC Medical Research Methodology* **12**: 99.

Holmes, C. C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression, *Bayesian Analysis* **1**(1): 145–168.

Ibrahim, J. G. and Chen, M.-H. (2000). Power prior distributions for regression models, *Statistical Science* **15**(1): 46–60. Available from: http://www.jstor.org/stable/2676676.

Karim, M. R. and Zeger, S. L. (1992). Generalized linear models with random effects; salamander mating revisited, *Biometrics* **48**(2): 631–644. Available from: http://www.jstor.org/stable/2532317.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data, *Biometrics* **38**(4): pp. 963–974. Available from: http://www.jstor.org/stable/2529876.

Lesaffre, E. and Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random effects model: an example, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **50**(3): 325–335.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**(1): pp. 13–22. Available from: http://www.jstor.org/stable/2336267.

Lin, X. and Breslow, N. E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion, *Journal of the American Statistical Association* **91**(435): pp. 1007–1016. Available from: http://www.jstor.org/stable/2291720.

Lindgren, F., Rue, H. and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(4): 423–498. Available from: http://dx.doi.org/10.1111/j.1467-9868.2011.00777.x.

Lu, G. and Ades, A. E. (2006). Assessing evidence inconsistency in mixed treatment comparisons, *Journal of the American Statistical Association* **101**(474): 447–459. Available from: http://dx.doi.org/10.1198/016214505000001302.

Lumley, T. (2002). Network meta-analysis for indirect treatment comparisons, *Statistics in Medicine* **21**(16): 2313–2324. Available from: http://dx.doi.org/10.1002/sim.1201.

Lunn, D. J., Thomas, A., Best, N. and Spiegelhalter, D. (2000). WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility, *Statistics and Computing* **10**(4): 325–337.

Lunn, D., Spiegelhalter, D., Thomas, A. and Best, N. (2009). The BUGS project: Evolution, critique and future directions, *Statistics in Medicine* **28**(25): 3049–3067. Available from: http://dx.doi.org/10.1002/sim.3680.

Martino, S., Akerkar, R. and Rue, H. (2011). Approximate Bayesian inference for survival models, *Scandinavian Journal of Statistics* **38**(3): 514–528. Available from: http://dx.doi.org/10.1111/j.1467-9469.2010.00715.x.

Martins, T. G., Simpson, D., Lindgren, F. and Rue, H. (2013). Bayesian computing with INLA: New features, *Computational Statistics & Data Analysis* **67**(0): 68 – 83. Available from: http://www.sciencedirect.com/science/article/pii/S0167947313001552.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, number 37 in *Monographs on Statistics and Applied Probability*, second edn, Chapman and Hall, New York.

Metropolis, N., Rosenbluth, A. W., Rosenbluth1, M. N., Teller, A. H. and Teller, E. (1953). Equation of state calculations by fast computing machines, *Journal of Chemical Physics*.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*, Springer.

Muff, S., Riebler, A., Held, L., Rue, H. and Saner, P. (2015). Bayesian analysis of measurement error models using integrated nested Laplace approximations, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **64**(2): 231–252. Available from: http://dx.doi.org/10.1111/rssc.12069.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models, *Journal of the Royal Statistical Society. Series A (General)* **135**(3): 370–384. Available from: http://www.jstor.org/stable/2344614.

Paul, M., Riebler, A., Bachmann, L. M., Rue, H. and Held, L. (2010). Bayesian bivariate meta-analysis of diagnostic test studies using integrated nested Laplace approximations, *Statistics in Medicine* **29**(12): 1325–1339. Available from: http://dx.doi.org/10.1002/sim.3858.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. and R Core Team (2015). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-120. Available from: http://CRAN.R-project.org/package=nlme.

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model, *Journal of Computational and Graphical Statistics* **4**(1): 12–35.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling, *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.

Popper, K. R. (1959). *The Logic of Scientific Discovery*, Hutchinson.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. Available from: http://www.R-project.org/.

Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature, *Stata Journal* **2**(1): 1–21(21). Available from: http://www.stata-journal.com/article.html?article=st0005.

Rao, C. R. (1958). Some statistical methods for comparison of growth curves, *Biometrics* **14**(1): pp. 1–17. Available from: http://www.jstor.org/stable/2527726.

Raudenbush, S. W., Yang, M.-L. and Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation, *Journal of Computational and Graphical Statistics* **9**(1): 141–157. Available from: http://amstat.tandfonline.com/doi/abs/10.1080/10618600.2000.10474870.

Riebler, A., Held, L. and Rue, H. (2012). Estimation and extrapolation of time trends in registry data - Borrowing strength from related populations, *The Annals of Applied Statistics* **6**(1): 304–333. Available from: http://dx.doi.org/10.1214/11-AOAS498.

Robert, C. and Casella, G. (2011). A short history of Markov chain Monte Carlo: Subjective recollections from incomplete data, *Statist. Sci.* **26**(1): 102–115. Available from: http://dx.doi.org/10.1214/10-STS351.

Roos, M. and Held, L. (2011). Sensitivity analysis in Bayesian generalized linear mixed models for binary data, *Bayesian Analysis* **6**(2): 259–278.

Roos, M., Martins, T. G., Held, L. and Rue, H. (2015). Sensitivity analysis for Bayesian hierarchical models, *Bayesian Analysis* **10**(2): 321–349. Available from: http://dx.doi.org/10.1214/14-BA909.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, Chapman & Hall/CRC Press, London.

Rue, H. and Martino, S. (2007). Approximate Bayesian inference for hierarchical Gaussian Markov random field models, *Journal of Statistical Planning and Inference* **137**(10): 3177 – 3192. Special Issue: Bayesian Inference for Stochastic Processes. Available from: http://www.sciencedirect.com/science/article/pii/S0378375807000845.

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion), *Journal of the Royal Statistical Society - Series B* **71**: 319–392. Available from: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2008.00700.x/full.

Rue, H., Martino, S., Lindgren, F., Simpson, D., Riebler, A. and Krainski, E. T. (2014). *INLA: Functions which allow to perform full Bayesian analysis of latent Gaussian models using Integrated Nested Laplace Approximaxion*.

Ruppert, D., Wand, M. and Carroll, R. (2009). Semiparametric regression during 2003–2007, *Electronic Journal of Statistics* **3**(1): 1193–1256.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge.

Scheffé, H. (1956). Alternative models for the analysis of variance, *Annals of Mathematical Statistics* **27**(2): 251–271. Available from: http://dx.doi.org/10.1214/aoms/1177728258.

Schrödle, B., Held, L., Riebler, A. and Danuser, J. (2011). Using INLA for the evaluation of veterinary surveillance data from Switzerland: A case study, *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **60**(2): 261–279.

Stan Development Team (2014). *Stan Modeling Language Users Guide and Reference Manual, Version 2.5.0*. Available from: http://mc-stan.org/.

Stiratelli, R., Laird, N. and Ware, J. H. (1984). Random-effects models for serial observations with binary response, *Biometrics* **40**(4): pp. 961–971. Available from: http://www.jstor.org/stable/2531147.

Stroud, A. H. and Secrest, D. (1966). *Gaussian Quadrature Formulas*, Englewood Cliffs : Prentice-Hall.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities, *Journal of the American Statistical Association* **81**(393): 82–86. Available from: http://www.jstor.org/stable/2287970.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*, fourth edn, Springer, New York. Available from: http://www.stats.ox.ac.uk/pub/MASS4.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*, Springer.

Verbeke, G. and Molenberghs, G. (2013). The gradient function as an exploratory goodness-of-fit assessment of the random-effects distribution in mixed models, *Biostatistics* **14**(3): 477–490. Available from: http://biostatistics.oxfordjournals.org/content/14/3/477.abstract.

Veroniki, A. A., Vasiliadis, H. S., Higgins, J. P. T. and Salanti, G. (2013). Evaluation of inconsistency in networks of interventions, *International Journal of Epidemiology* **42**(1): 332–345. Available from: http://ije.oxfordjournals.org/content/42/1/332.abstract.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method, *Biometrika* **61**(3): pp. 439–447. Available from: http://www.jstor.org/stable/2334725.

Wishart, J. (1938). Growth-rate determinations in nutrition studies with the bacon pig, and their analysis, *Biometrika* **30**(1/2): pp. 16–28. Available from: http://www.jstor.org/stable/2332221.

Zhang, H., Lu, N., Feng, C., Thurston, S. W., Xia, Y., Zhu, L. and Tu, X. M. (2011). On fitting generalized linear mixed-effects models for binary responses using different statistical packages, *Statistics in Medicine* **30**(20): 2562–2572. Available from: http://dx.doi.org/10.1002/sim.4265.

PAPER I

CD8 counts and CD4/CD8 ratio independently predict CD4 response in drug naive and in patients on cART

Rafael Sauter, Ruizhu Huang, Bruno Ledergerber, Manuel Battegay, Enos Bernasconi, Matthias Cavassini, Hansjakob Furrer, Matthias Hoffmann, Mathieu Rougemont, Huldrych F. Günthard, Leonhard Held & the Swiss HIV cohort study

Paper submitted to *Journal of Acquired Immune Deficiency Syndromes*.

CD8 counts and CD4/CD8 ratio independently predict CD4 response in drug naive and in patients on cART

Rafael SAUTER[a,*], Ruizhu HUANG[a], Bruno LEDERGERBER[b], Manuel BATTEGAY[d], Enos BERNASCONI[e], Matthias CAVASSINI[f], Hansjakob FURRER[g], Matthias HOFFMANN[h], Mathieu ROUGEMONT[i], Huldrych F GÜNTHARD[b,c], Leonhard HELD[a], and the Swiss HIV cohort study[1]

[a]*Epidemiology, Biostatistics and Prevention Institute, University of Zurich*
[b]*Division of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich*
[c]*Institute of Medical Virology, University of Zurich*
[d]*Division of Infectious Diseases and Hospital Epidemiology, University Hospital Basel*
[e]*Division of Infectious Diseases, Regional Hospital Lugano*
[f]*Division of Infectious Diseases, University Hospital Lausanne*
[g]*Department of Infectious Diseases, Bern University Hospital, University of Bern*
[h]*Division of Infectious Diseases and Hospital Epidemiology, Cantonal Hospital St.Gallen*
[i]*Division of Infectious Diseases, University Hospital Geneva*

Background: *Plasma HIV viral load is related to declining CD4 lymphocytes. The extent to which CD8 cells, in addition to RNA viral load, predict the depletion of CD4 cells is not well characterized so far. We examine if CD8 cell counts is a prognostic factor for CD4 cell counts during an HIV infection.*

Methods: *A longitudinal analysis is conducted using data from the Swiss HIV cohort study collected between January 2000 and October 2014. Linear mixed regression models were applied to observations from HIV-1 infected treatment naive patients (NAIVE) and cART treated patients to predict the short-term evolution of CD4 cell counts. For each subgroup it was quantified to which extent CD8 cell counts or CD4/CD8 ratios are prognostic factors for disease progression.*

Results: *In both subgroups, 2'500 NAIVE and 8'902 cART patients, past CD4 cells are positively ($p<0.0001$) and past viral load is negatively ($p<0.0001$) associated with the outcome. Including additionally past CD8 cell counts improves the fit significantly ($p<0.0001$) and increases the marginal explained variation 31.7% to 40.7% for the NAIVE and from 44.1% to 50.7% for the cART group. The past CD4/CD8 ratio (instead of the past CD8 level) is pos-*

*Corresponding author: rafael.sauter@uzh.ch
[1]Members of the Swiss HIV Cohort Study Group are provided in the Acknowledgments.

itively associated with the outcome, increasing the explained variation further to 41.8% for NAIVE and 51.9% for cART.

Conclusions: *CD8 lymphocytes contain essential information on HIV-1 disease progression. Incorporating CD4/CD8 ratio in combination with CD4 counts is more informative among NAIVE as well as for cART patients, compared to the use of the level of both lymphocyte subtypes.*

1. Introduction

An untreated HIV-1 infection is characterized by declining CD4 target cells which is associated with the viral load level. Over time, viral load levels in general tend to increase and CD4 levels continue to decline with subsequent cellular immunodeficiency leading to AIDS and ultimately death [1, 2]. Successful antiretroviral treatment (ART) results in sustained suppression of HIV-1 plasma RNA levels below the detection limit of currently available assays. Today's combined antiretroviral therapy (cART) is a combination of at least three different substances consisting of a non-nucleoside reverse transcriptase, a boosted protease or an integrase inhibitor with a combined two drug nucleoside/nucleotide backbone [3].

Past research based on randomized trials and cohort studies mainly focused on the HIV-1 plasma RNA load and CD4 cell count interactions over time [4, 5, 6] and the restoration of the CD4 cell counts [7, 8, 9]. However, already in the early times of HIV research it was suggested to include additional immune-activation measures such as CD8 lymphocyte cell counts, CD4/CD8 ratios or CD4 and CD8 percentages [10, 11, 12, 13]. A negative correlation between changes in CD4 and CD8 cell counts during an intensification of the antiretroviral therapy was reported [14]. In the Swiss HIV cohort study (SHCS) larger changes in CD4 cell counts were found to be negatively associated with CD8 cell counts measured at baseline for HIV-1 patients receiving antiretroviral therapy [15]. For HIV-1 infected treatment naive patients, CD8 counts increase while CD4 counts decline [16] but only viral load and CD4 counts and not CD8 cell counts, were considered to be the most relevant predictors for disease progression [17]. Time to normalisation of the CD4/CD8 ratio, defined as two subsequent measurements with a ratio above between one and 1.2 was found to be negatively associated

with its baseline value [18, 19] but only a minority of HIV-1 infected individuals under antiretroviral therapy normalize their CD4/CD8 ratio [18, 19, 20], in particular if treatment was started at low CD4 counts [15, 21, 3]. Low CD4/CD8 ratios were also found to be associated with increased morbidity and mortality of non-AIDS related death causes [19, 22, 23].

These studies all hint towards a possibly important role of CD8 cell counts during an HIV infection. However, up to now an analysis of the time-dependent relationship between changing CD8 and CD4 lymphocytes based on a cohort study is lacking. Furthermore, there is a large inter-patient variation in disease progression, in CD4 recovery under therapy and in CD4/CD8 normalization, depending on a multitude of factors such as viral and host factors [24, 25, 26]. Here, by taking patient-specific variation into account, we examined whether past CD8 cell counts contain additional information to determine future CD4 cell counts and investigated this effect separately, for treatment naive individuals and for patients receiving cART.

2. Methods

2.1. Study population

The SHCS [27], established in 1988, includes HIV-1 infected persons older than 18 years, living in Switzerland. The SHCS schedules regular follow-up visits every six months, while the common clinical follow-up interval is three months, at which CD4 and CD8 lymphocyte cell counts and plasma HIV-1 viral load are measured. The lymphocyte cell counts per μL blood were measured by flow cytometry. Since the year 2000 all assays used for HIV-1 RNA detection had a detection limit of 50 copies per mL or lower. For this study, the RNA detection limit was set at 50 RNA copies/mL of plasma, independent of the applied assays and all values below this limit, or without detection, were set to 25 copies/mL. Data were extracted from the October 2014 update of the SHCS database. Observations prior to the year 2000 were excluded in order to guarantee comparable assay technology used to measure plasma RNA load and that an established cART was available to all patients. We extracted from the database 280'554 lymphocyte cell counts and 325'984 RNA measurements obtained from 11'899 patients.

The study population was divided into two subgroups, one covers all observations obtained from patients with an untreated HIV-1 infection (NAIVE), observed as long

as they did not start cART. The second group includes observations from patients receiving available standard cART. Accordingly the same patient may be included in both groups, which is the case for 1'797 patients or 71.9% of the NAIVE study population. Lymphocytes and RNA are not always measured at the same time, so we matched results of the two laboratory analyses if the time difference was less than eight days and the date of the RNA analysis was kept Observed lymphocyte cell counts for which no RNA measurement was available were omitted. If a patient quit or interrupted cART therapy, all follow-up observations were omitted, independent of a likely therapy resumption. As it was shown that Hepatitis C co-infection influences CD4 cell counts, all patients with indetermined HCV status were excluded [28]. Moreover, all observations with a follow-up time between two subsequent measurements of more than twelve months were excluded, as the information of past lymphocyte cell counts and past RNA measurements for future CD4 counts was assumed to diminish over time. If a patient met all of the above inclusion criteria he additionally had to have at least three measurements of CD4 and CD8 cell counts as well as RNA blood viral load, observed at three different occasions. The selection of the study population according to these eligibility criteria is described in Figure 1.

2.2. Statistical methods and analysis

The hypothesis that past CD4/CD8 ratios predict current CD4 cell counts was examined by linear mixed regression models for longitudinal data [29] for each patient subgroup. The outcome in each model is the square root transformed CD4 cell count [4, 30] observed at the current follow-up visit. We estimated for both patient groups three models for which we included different combinations of suitably transformed CD4 and CD8 cell counts, observed at the preceding follow-up visit, as predictors. In this way the influence of CD4 and CD8 lymphocytes, which are lagged by one follow-up visit, on the outcome is examined. The lag between two subsequent follow-up visits corresponds on average to a time period of three months. From here on predictors are called lagged if, relative to the observed CD4 cell count outcome, they were observed at the preceding follow-up visits. In the first model formulation (M1) we included the lagged square root transformed CD4 cell counts and the lagged $\log_{10}$ transformed RNA measurement as predictors, both observed at the preceding of two subsequent follow-

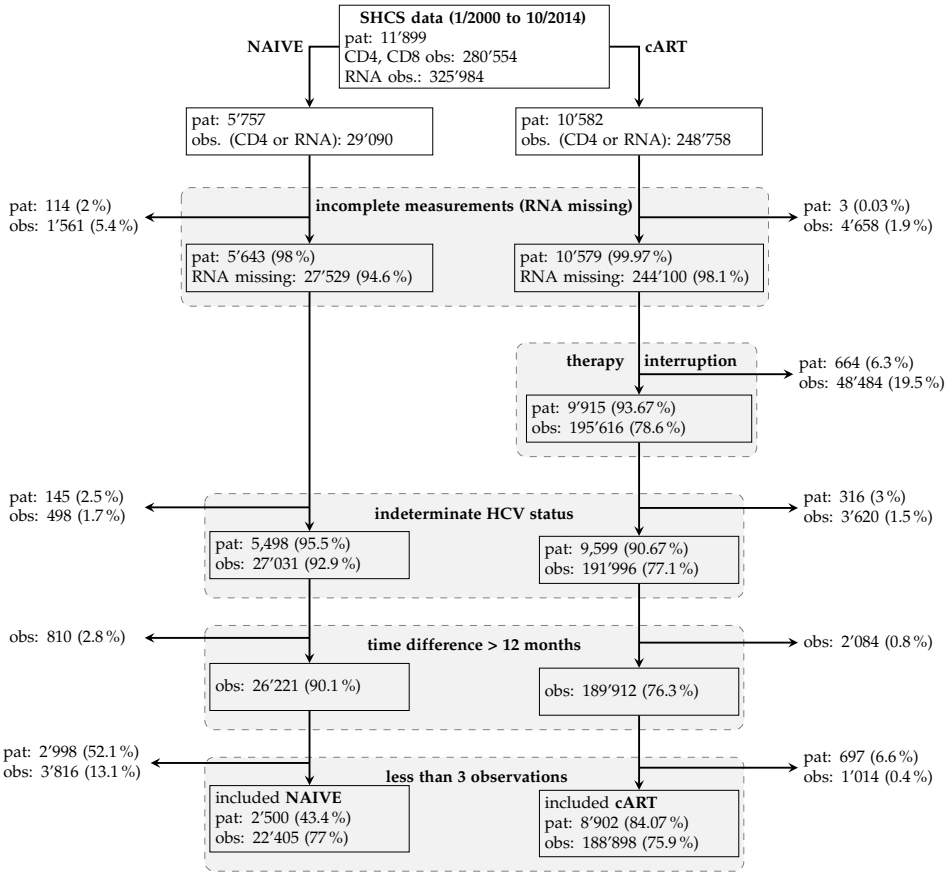

Figure 1: SHCS data and study population with number of observations (obs.) and number of patients (pat.). Percentages (%) are based on the initial NAIVE or cART data given on the second level in the diagram.

up visit. In the second model (M2) we added the lagged square root transformed CD8 cell counts as additional predictor. In the third approach (M3) we replaced the lagged CD8 cell counts by the natural log transformed, lagged CD4/CD8 ratio.

In order to assess which of the three model is most suitable for each patient group we compared the marginal R^2 [31, 32], which is the proportion of explained variation by the predictors as proportion of the overall variation. Additionally the models were compared by a version of the Bayesian information criterion (BIC), modified for linear mixed models [33]. The modified BIC penalizes the inclusion of model parameters and lower values indicate that the corresponding model captures more information and thus should be preferred.

Time was set to zero at cohort entry for the NAIVE and at therapy initiation for the cART subgroup. The time scale was standardised to three month intervals, as this corresponds to the common clinical follow-up period in the SHCS. For the cART group time since therapy start was square root transformed [34], as this allowed to capture the sharp increase in CD4 cell counts after therapy initiation [35]. We also included AIDS (yes, no) and age, which are time-dependent, the transmission group (transmission) and the status of a hepatitis C co-infection (HCV), both observed at baseline, as predictors in all models. The probable HIV transmission [36] is a categorical predictor with six groups: homosexual men (MSM), male and female intravenous drug users (IDU-male, IDU-female), heterosexual males and females (HET-male, HET-female) and a group for which the transmission path is not further specified (other). The HCV status has three categories: HCV negative, patients with inactive and patients with replicating (active) HCV. For the cART patient group we additionally included the time period prior to cART during which an individual was receiving mono or dual regimens with nucleoside reverse-transcriptase inhibitors (NRTI).

In order to address patient-specific heterogeneity at baseline we included a random intercept in each model and heterogeneity between patients in the CD4 cell time course is taken into account by a random slope, which allows for patient-specific deviations from the average time course [4]. The random effect structure is the same for both subgroups and all three model specifications. An analysis of the model residuals did not provide evidence for any longitudinal structure or other dependency. All statistical

Baseline characteristics	**NAIVE (n=2'500)**		**cART (n=8'902)**	
CD4 at baseline	491	(376 , 670)	367	(231 , 539)
CD8 at baseline	904	(654 , 1265)	846	(601 , 1185)
CD4/CD8 at baseline	0.55	(0.38 , 0.8)	0.41	(0.25 , 0.67)
log(RNA) at baseline	4.23	(3.58 , 4.8)	1.95	(1.40 , 3.11)
AIDS at baseline	46	(1.8%)	1843	(20.7%)
age at baseline	36	(30 , 42)	39	(33 , 46)
transmission				
MSM	1181	(47.2%)	3723	(41.8%)
IDU-male	208	(8.3%)	841	(9.4%)
IDU-female	94	(3.8%)	423	(4.8%)
HET-male	415	(16.6%)	1602	(18.0%)
HET-female	506	(20.2%)	1892	(21.3%)
other	96	(3.8%)	421	(4.7%)
HCV				
HCV negative	2001	(80.0%)	7056	(79.3%)
HCV inactive	109	(4.4%)	367	(4.1%)
HCV active	390	(15.6%)	1479	(16.6%)
Longitudinal characteristics	**NAIVE (n=22'405)**		**cART (n=188'898)**	
CD4	457	(346 , 609)	497	(338 , 688)
CD8	952	(694 , 1300)	796	(573 , 1096)
CD4/CD8	0.48	(0.34 , 0.69)	0.62	(0.40 , 0.92)
log(RNA)	4.13	(3.50 , 4.66)	1.40	(1.40 , 1.40)
follow-up visits per patient	7	(4 , 12)	17	(9 , 31)
follow-up time (months)	3.5	(3.0 , 5.6)	3.3	(2.8 , 5.3)

Table 1: Patient characteristics for baseline (upper part) and longitudinal (lower part) characteristics of the study population for both subgroups. Data are patient numbers (with %) for discrete and median (with the first and third quartile) for continuous predictors. Lymphocyte cell counts and ratio are untransformed, the RNA is $\log_{10}$-transformed.

analyses and data preparation were done in R version 3.2.3 (2015-12-10). All linear mixed models were fitted with the software package `nlme` [37] version 3.1-119.

3. Results

3.1. Patient characteristics

According to the above inclusion criteria 2'500 patients were eligible in the NAIVE subgroup with 22'405 observations, while 8'902 patients were included in the cART subgroup with 188'898 observations (see Figure 1). The patient characteristics for both subgroups are shown in Table 1. The upper part in Table 1 shows the patient characteristics for predictors observed at baseline and the lower part for time-dependent, longitudinal predictors based on all included observations.

Patients in the NAIVE, compared to the cART subgroup, had lower CD4 cell counts (NAIVE 457, cART 497), higher CD8 cell counts (NAIVE 952, cART 796), a lower CD4/CD8 ratio (NAIVE 0.48, cART 0.62), and higher log-transformed RNA levels (NAIVE 4.13, cART 1.40). The cART group had more follow-up visits per patient (NAIVE 46, cART 1843), as usually most of the patients start cART rather quickly and continuously stay under therapy for the remaining observation time, thus patients are also older at baseline in the treated group (NAIVE 36, cART 39). The proportion of patients having AIDS was higher in the cART group (NAIVE 4.9%, cART 24.1%). All other characteristics (follow-up time, transmission, HCV) were similar for both subgroups.

3.2. Model comparison and predictive value of CD8 lymphocytes

The first two columns in Table 2 give the marginal R^2 for the three models and for each patient group. For the NAIVE subgroup the marginal R^2 increases from 31.7% for M1, to 40.7% for M2 after inclusion of the lagged CD8 cell counts. For cART the increase is from 44.1% to 50.7%. Including the lagged CD4/CD8 ratio instead of the lagged CD8 cell counts, increases the R^2 measure even more to 41.8% for the NAIVE and to 51.9% for the cART group. The modified BIC for the three models and each patient group is shown in the last two columns of Table 2. The modified BIC is decreasing from model M1 to M3 for both patient groups. Thus, consistent with the marginal R^2, also the information criterion prefers the model which includes the CD4/CD8 ratio over the other two models.

	marginal R^2		**modified BIC**	
	NAIVE	**cART**	**NAIVE**	**cART**
M1: CD4 + RNA	31.7%	44.1%	95'594	865'332
M2: CD4 + CD8 + RNA	40.7%	50.7%	95'060	862'400
M3: CD4 + CD4/CD8 + RNA	41.8%	51.9%	94'962	861'192

Table 2: Marginal R^2 and modified BIC for both patient subgroups and three models including different lagged longitudinal predictors. All models additionally include an intercept and were adjusted for AIDS, age, transmission, HCV and for time since cohort entry for NAIVE and time since therapy start plus NRTI at baseline for cART (see also Table 3).

Figure 2 illustrates the relationship between the lagged lymphocyte predictors, observed at the first follow-up visit and the outcome, observed at the second of two subsequent follow-up visits. Figure 2 depicts the relationship for the population mean in the NAIVE (left) and cART (right) group. Square root transformed lymphocyte cell counts were back transformed to absolute cell count values and the log-transformation of the CD4/CD8 ratio was also reversed. The lower and upper limits for the predictors in Figure 2 were set to the 2.5% and 97.5% quantiles of all observation (0.13 to 1.8 for the lagged CD4/CD8 ratio and 114 to 1184 for the lagged CD4). The contour lines of the plots in the first two rows indicate the predicted CD4 cell count at the next follow-up visit, for a given combination of lagged CD4/CD8 ratio and lagged CD4 cell count. All other predictors in model M2 and M3 were set to a constant value: time since cohort entry for NAIVE or since therapy start for cART was set to three months, the age to 40 years, we assumed a median viral load equal to 4.1 (corresponding to 13'380 copies per *ml* blood) for NAIVE and a suppressed RNA level for cART (equal to a viral load of 1.4 or 25 copies per *ml* blood), which occurs for 76.4% of the observations in this group. The categorical predictors, transmission and HCV, were set to the reference category and NRTI duration was set to zero.

Figure 2 demonstrates that for model M2 and M3 the relationship between the lymphocyte predictors and the outcome is similar for untreated and cART patients. However, the two models give considerably different predictions. The plots for cART and NAIVE based on model M2 (first row) show converging contour lines in the upper left corner. This leads to almost vertical contour lines for Model M2, if the CD4/CD8 ratio is below 0.25, implying a CD8 count which is four times larger than the CD4 cell count. The vertical contour lines mean that in this range the CD4 lymphocyte level has, according to Model M2, no more impact on the prediction of CD4 at the next follow-up visit. Model M3 in contrast, which instead of the CD8 cell counts includes the CD4/CD8 ratio as predictor, does not have this feature of quickly converging contour lines (second row). In model M3 the level of CD4 lymphocytes is still an important predictor, even if the CD4/CD8 ratio is below 0.25.

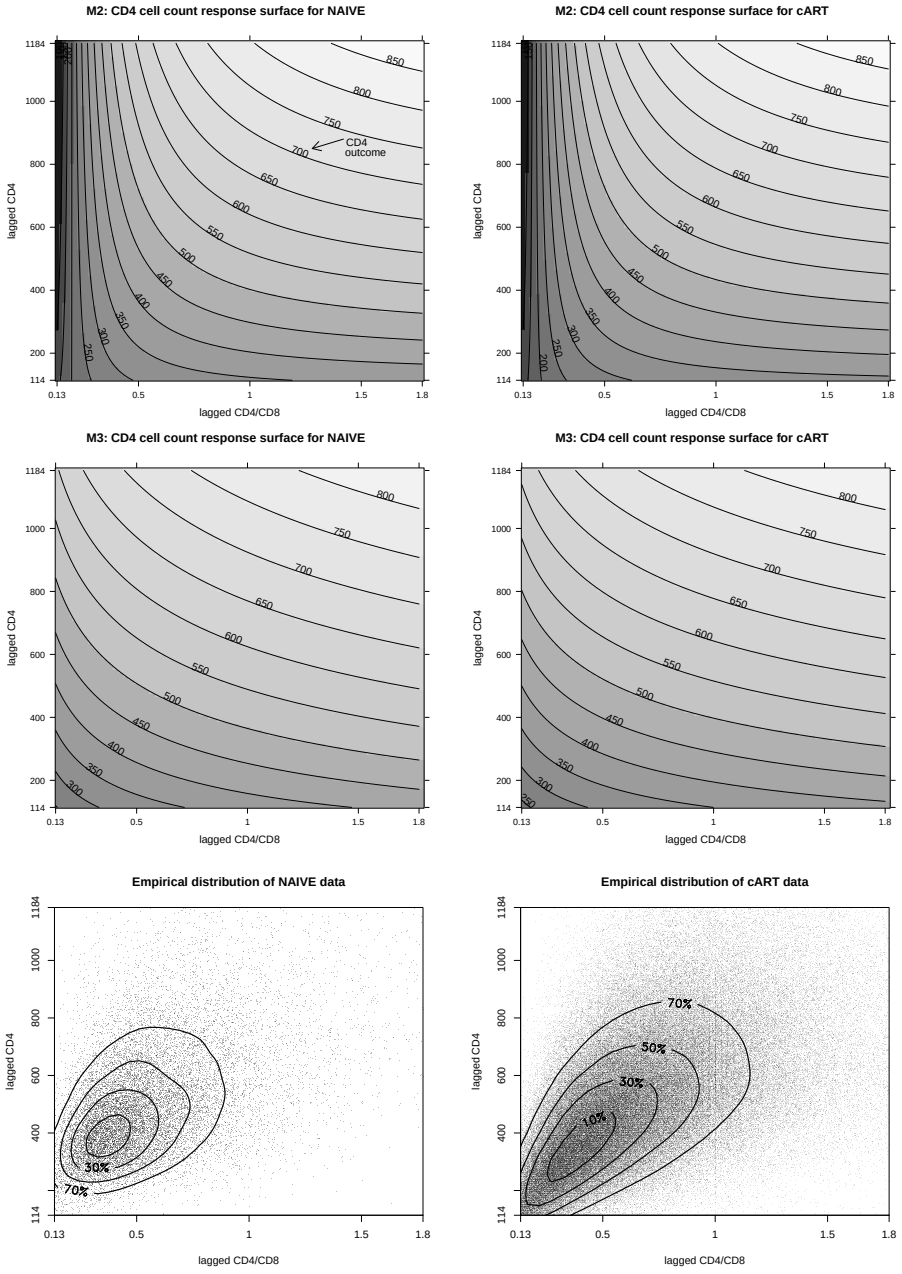

Figure 2: Relation between lagged CD4/CD8 ratio predictor (abscissa), lagged CD4 cell counts predictor (ordinate) and the predicted CD4 cell count outcome (contour lines) for the population mean in NAIVE (left) and cART (right) subgroups in model M2 (top) and M3 (middle) and the empirical distribution of the data (bottom) for which the contour lines indicate the location of 70%, 50%, 30% and 10% of the data.

The bottom row in Figure 2 shows the empirical bivariate distribution of CD4 cell counts and the CD4/CD8 ratio for NAIVE and cART patients. The contour lines indicate where 70%, 50% and 10% of the observations are located. For the NAIVE subgroup 8.3% and for cART 9.7% of the observed data are below a CD4/CD8 ratio equal to 0.25. The empirical distribution in Figure 2 underpins the fact, that the differences between model M2 and M3 are relevant, as there is a substantial proportion of observations situated in the bottom left corner, where the models differ most.

3.3. Estimates of regression coefficients

The upper part in Table 3 shows the regression coefficients together with a 95% confidence interval (CI) for the intercept and the longitudinal predictors, based on model M3 only which according to the criteria in Table 2 and Figure 2 is the preferred model. The intercept represents the population mean of the square root transformed CD4 cell counts at baseline (18.02 NAIVE, 15.56 cART). The regression coefficient for time reflects the population mean course of the outcome since cohort entry for the NAIVE (-0.10) and since therapy start for cART (0.34). These estimates are in agreement with the established finding that for untreated HIV infected patients the CD4 count declines, while an antiretroviral therapy causes an increase. The lagged square root CD4 cell count, observed at the preceding follow-up visit, is positively associated with the outcome and has a similar coefficient for both subgroups (0.36 NAIVE and 0.39 cART). The negative coefficient for the plasma viral load (-0.46 NAIVE and -0.27 cART) is in line with the acknowledged decline of CD4 cell counts for higher viral loads. The lagged log-transformed CD4/CD8 ratio is positively associated with the outcome for both subgroups (1.79 NAIVE and 1.65 cART). As the CD4/CD8 ratio is log-transformed, this implies a positive effect if CD4 exceeds the CD8 cell count, but a negative effect if the ratio is below one. A patient with an imbalanced immune system, for whom the CD4 is below the CD8 cell count, has a lower prediction for the CD4 cells at the next follow-up visit, compared to a patient with equal CD4 count but with a higher CD4/CD8 ratio.

The regression coefficients for the remaining predictors are shown in the lower part of Table 3. In contrast to the predictors in the upper part, these predictors are not time-dependent and have an effect on the predicted CD4 cell count level only. However, they do not imply differences in the decline or relapse of the outcome, as they do not include

	NAIVE			cART		
	coef.	95% CI	p-value	coef.	95% CI	p-value
intercept	18.02	17.43 to 18.61	< 0.0001	15.56	15.32 to 15.80	< 0.0001
time	-0.10	-0.11 to -0.10	< 0.0001			< 0.0001
square root time			< 0.0001	0.34	0.32 to 0.36	< 0.0001
square root CD4	0.36	0.35 to 0.38	< 0.0001	0.39	0.39 to 0.40	< 0.0001
log RNA	-0.46	-0.53 to -0.40	< 0.0001	-0.27	-0.29 to -0.25	< 0.0001
log CD4/CD8	1.79	1.66 to 1.93	< 0.0001	1.65	1.60 to 1.70	< 0.0001
AIDS at follow-up visit	-0.88	-1.28 to -0.48	< 0.0001	-1.01	-1.12 to -0.90	< 0.0001
age at follow-up visit	-0.01	-0.02 to 0.00	0.0018	-0.03	-0.03 to -0.02	< 0.0001
NRTI at baseline				-0.02	-0.02 to -0.01	0.00013
transmission			0.01			< 0.0001
MSM (reference)	0.00			0.00		
IDU-male	-0.14	-0.61 to 0.32		-1.13	-1.39 to -0.88	
IDU-female	-0.12	-0.71 to 0.46		-0.75	-1.05 to -0.44	
HET-male	-0.32	-0.60 to -0.05		-0.64	-0.79 to -0.49	
HET-female	-0.40	-0.66 to -0.15		-0.47	-0.61 to -0.32	
other	-0.04	-0.55 to 0.46		-0.34	-0.60 to -0.08	
HCV			0.47			< 0.0001
negative (reference)	0.00			0.00		
inactive	-0.08	-0.59 to 0.44		0.04	-0.26 to 0.34	
active	-0.13	-0.49 to 0.23		-0.44	-0.64 to -0.23	

Table 3: Coefficient estimates (coef.) with 95% confidence intervals (CI) and p-values for model M3 and for both patient subgroups (NAIVE and cART). Upper part shows longitudinal predictors and the lower part time constant predictors for AIDS and age at follow-up visit, NRTI at baseline, transmission and HCV.

any interactions with time dependent longitudinal predictors reported in the upper part of Table 3. If a patient had AIDS at the time of observation, the CD4 cell count will be lower. Also the patients age and the duration of NRTI prior to cART has a negative level effect on the outcome. The estimates for the transmission group provides evidence that the heterosexual transmission categories (HET-male, HET-female) are different from the homosexual men reference category (MSM) in the NAIVE group as the upper bound of the 95% CI is negative. For the cART group the 95% CI's imply that, compared to the reference category (MSM), all other transmission categories have a lower CD4 cell count level. For the HCV predictor in the NAIVE group there is no evidence for a difference between the three categories. However, in the cART group the corresponding p-value is small, implying that at least one HCV category has a different CD4 level compared to the HCV negative reference category, which probably concerns the active HCV category.The Supplementary Material provides results for model M1 and M2, as

well as additional information.

4. Discussion

Based on a large dataset from the Swiss HIV cohort study we have shown that CD8 cell counts contain crucial predictive information for the HIV disease progression in drug naive patients and as well for the CD4 cell count recovery in patients receiving cART. Both lymphocyte cell subtypes, CD4 and CD8, as well as the RNA viral load, have been identified as important prognostic factors for the CD4 cell count over the next months. We could show that the model which includes the lagged levels of CD4 and CD8 as two separate predictors was inferior compared to an approach which included the lagged CD4/CD8 ratio instead of the CD8 level. These findings persisted, also if we applied alternative transformations to the lagged lymphocyte predictors, such as the log instead of the square root transformation. The CD4/CD8 ratio can be interpreted as a measure for the imbalance of the patient's immune system which captures essential information, additional to the cell count levels of both lymphocyte subtypes. The CD8 cell count is a marker for immune activation, which has been found to be an important factor for disease progression [38, 39].

The relationship between the lagged CD4 cell count predictor, the lagged CD4/CD8 ratio predictor and the CD4 cell count at the next follow-up visit were found to be surprisingly similar for untreated patients and patients under a cART. This was illustrated by the response surface for the CD4 cell counts in Figure 2 and the comparable estimates between NAIVE and cART for the lymphocyte predictors shown in Table 3.

Furthermore we could quantify the relative importance of the two lymphocyte subtypes and the viral load as prognostic factors. It was previously found that the viral load only explains a low proportion (below 10%) of the variation for changes in CD4 cell counts over one year [6]. Such a statement crucially depends on the outcome and time period for which the prediction is made. We found that even for NAIVE patients lagged CD8 cell counts are more important than lagged viral loads in explaining the variation in future CD4 cell counts. If the lagged RNA predictor is omitted then the marginal R^2 reduces from 31.7% to 27.2% in the approach which includes lagged CD4 as predictor. For the cART group the marginal R^2 is even increasing for the same model

formulation, if the viral load predictor is omitted. This reflects the fact that the association between CD4 cell counts and viral load is fundamentally different for treatment naive and patients under therapy. The improved marginal R^2 and also the difference in the regression coefficients of the lagged RNA for the NAIVE and cART group, supports the argument that although the viral load is a relevant predictor for future CD4 cell counts, it explains perhaps only a small proportion of the observed variation. As the plasma RNA load is often suppressed below the assay detection limit under cART we also examined the possible influence of recurrent detectable plasma RNA viral loads ("blips"), but no evidence for an additional effect was found.

By choosing a linear mixed model for the longitudinal SHCS data we could incorporate patient specific-variation for the CD4 cell count at baseline and for its patient-specific time course. The applied model simplified the observation pattern, ignoring interval censored data and assumed regular follow-up times, although in the SHCS data there is notable variation for the time between two follow-up visits (see Supplementary Material).

Of course, causal pathways between the quantities of interest, CD4, CD8 and RNA, are inherently difficult to assess and not possible to estimate without a minimal set of assumptions. Nevertheless, we have derived a statistical model, which allowed to address the clinically important question of how the patients immune system will evolve given the current state. It would be of interest to examine the sensitivity of the results with respect to the underlying assumptions. For example an alternative analysis would incorporate the underlying rather than the observed lymphocyte and viral load measurements as predictors [40].

With this statistical analysis of a large HIV cohort we could confirm the findings by other studies, which attributed an important role to CD8 cells for describing the HIV-1 disease progression. We were able to show that the CD4/CD8 ratio is an important time-dependent prognostic factor, both for treatment naive and cART-treated patients. The SHCS data could be used for further investigations concerning the role of CD8 during an HIV-1 infection, which could reveal more details about the connections between HIV-1 disease progression and variation for the time needed to normalize the CD4/CD8 ratio. Elaborating such associations in more detail, also for interactions of

other risk factors [41, 42], would support the understanding of the mechanisms leading to the high variation in different individual immune responses.

Acknowledgements

This study was supported by the Swiss National Science Foundation (SNF grant #33CS30-134277), and the Swiss HIV Cohort Study (SHCS project 725). The SHCS data are gathered by the Five Swiss University Hospitals, two Cantonal Hospitals, 15 affiliated hospitals and 36 private physicians (listed in `http://www.shcs.ch/180-health-care-providers`). Members of the Swiss HIV Cohort Study are: Aubert V, Battegay M, Bernasconi E, Böni J, Braun DL, Bucher HC, Burton-Jeangros C, Calmy A, Cavassini M, Dollenmaier G, Egger M, Elzi L, Fehr J, Fellay J, Furrer H (Chairman of the Clinical and Laboratory Committee), Fux CA, Gorgievski M, Günthard H (President of the SHCS), Haerry D (deputy of "Positive Council"), Hasse B, Hirsch HH, Hoffmann M, Hösli I, Kahlert C, Kaiser L, Keiser O, Klimkait T, Kouyos R, Kovari H, Ledergerber B, Martinetti G, Martinez de Tejada B, Marzolini C, Metzner K, Müller N, Nadal D, Nicca D, Pantaleo G, Rauch A (Chairman of the Scientific Board), Regenass S, Rudin C (Chairman of the Mother & Child Substudy), Schöni-Affolter F (Head of Data Centre), Schmid P, Speck R, Stöckle M, Tarr P, Trkola A, Vernazza P, Weber R, Yerly S. (version June 2015).

References

[1] J. W. Mellors, L. A. Kingsley, C. R. Rinaldo, J. A. Todd, B. S. Hoo, R. P. Kokka, et al., Quantitation of HIV-1 RNA in plasma predicts outcome after seroconversion, Annals of Internal Medicine 122 (8) (1995) 573–579.

[2] J. W. Mellors, C. R. Rinaldo, P. Gupta, R. M. White, J. A. Todd, L. A. Kingsley, Prognosis in HIV-1 infection predicted by the quantity of virus in plasma, Science 272 (5265) (1996) 1167–1170.

[3] H. Günthard, J. Aberg, J. Eron, J. Hoy, A. Telenti, C. Benson, et al., Antiretroviral treatment of adult hiv infection: 2014 recommendations of the international antiviral society–usa panel, Journal of the American Medical Association 312 (4) (2014) 410–425.

[4] J. W. Boscardin, J. M. Taylor, N. Law, Longitudinal models for AIDS marker data, Statistical Methods In Medical Research 7 (1998) 13–27.

[5] R. Thiébaut, H. Jacqmin-Gadda, A. Babiker, D. Commenges, Joint modelling of bivariate longitudinal data with informative dropout and left-censoring, with application to the evolution of CD4+ cell count and HIV RNA viral load in response to treatment of HIV infection, Statistics in Medicine 24 (2005) 65–82.

[6] B. Rodríguez, A. Sethi, V. Cheruvu, W. Mackay, R. Bosch, M. Kitahata, et al, Predictive value of plasma HIV RNA level on rate of CD4 t-cell decline in untreated HIV infection, Journal of the American Medical Association 296 (12) (2006) 1498–1506.

[7] G. R. Kaufmann, H. Furrer, B. Ledergerber, L. Perrin, M. Opravil, P. Vernazza, et et., Characteristics, determinants, and clinical relevance of CD4 T cell recovery to <500 cells/µL in HIV type 1- infected individuals receiving potent antiretroviral therapy, Clinical Infectious Diseases 41 (3) (2005) 361–372.

[8] M. Battegay, R. Nüesch, B. Hirschel, G. R. Kaufmann, Immunological recovery and antiretroviral therapy in HIV-1 infection, The Lancet Infectious Diseases 6 (5) (2006) 280 – 287.

[9] N. Khanna, M. Opravil, H. Furrer, M. Cavassini, P. Vernazza, E. Bernasconi, et al., CD4+ T cell count recovery in HIV type 1 - infected patients is independent of class of antiretroviral therapy, Clinical Infectious Diseases 47 (8) (2008) 1093–1101.

[10] J. M. G. Taylor, J. L. Fahey, R. Detels, J. V. Giorgi, CD4 percentage, CD4 number, and CD4: CD8 ratio in HIV infection: which to choose and how to use, Journal of Acquired Immune Deficiency Syndromes 2 (2) (1989) 114–124.

[11] J. V. Giorgi, Z. Liu, L. E. Hultin, W. G. Cumberland, K. Hennessey, R. Detels, Elevated levels of CD38+ CD8+ T cells in HIV infection add to the prognostic value of low CD4+ T cell levels: results of 6 years of follow-up, Journal of Acquired Immune Deficiency Syndromes 6 (8) (1993) 904–912.

[12] J. M. Benito, M. López, S. Lozano, P. Martinez, J. González-Lahoz, V. Soriano, CD38 expression on CD81 T lymphocytes as a marker of residual virus replication in

chronically HIV-infected patients receiving antiretroviral therapy, AIDS Research and Human Retroviruses 20 (2) (2004) 227–233.

[13] J. W. T. Cohen Stuart, M. D. Hazebergh, D. Hamann, S. A. Otto, J. C. C. Borleffs, F. Miedeman, et al., The dominant source of CD4+ and CD8+ T-cell activation in HIV infection is antigenic stimulation, Journal of Acquired Immune Deficiency Syndromes 25 (3) (2000) 203–211.

[14] M. A. Kolber, M. O. Saenz, T. J. Tanner, K. L. Arheart, S. Pahwa, H. Liu, Intensification of a suppressive HAART regimen increases CD4 counts and decreases CD8+ T-cell activation, Clinical Immunology 126 (3) (2008) 315 – 321.

[15] G. Kaufmann, L. Perrin, G. Pantaleo, M. Opravil, H. Furrer, A. Telenti, et al., CD4 T-lymphocyte recovery in individuals with advanced HIV-1 infection receiving potent antiretroviral therapy for 4 years: The Swiss HIV cohort study, Archives of Internal Medicine 163 (18) (2003) 2187–2195.

[16] J. B. Margolick, A. Munoz, A. D. Donnenberg, L. P. Park, N. Galai, J. V. Giorgi, et al., Failure of T-cell homeostasis preceding AIDS in HIV-1 infection, Nature Medicine 1 (7) (1995) 674–680.

[17] J. W. Mellors, A. Munoz, J. V. Giorgi, J. B. Margolick, C. J. Tassoni, P. Gupta, et al., Plasma viral load and CD4+ lymphocytes as prognostic markers of HIV-1 infection, Annals of Internal Medicine 126 (12) (1997) 946–954.

[18] W. Tinago, E. Coghlan, A. Macken, J. McAndrews, B. Doak, C. Prior-Fuller, et al., Clinical, immunological and treatment - related factors associated with normalised CD4+/CD8+ T-cell ratio: effect of naïve and memory T-cell subsets, PLoS ONE 9 (5) (2014).

[19] C. Mussini, P. Lorenzini, A. Cozzi-Lepri, G. Lapadula, G. Marchetti, E. Nicastri, et al., CD4/CD8 ratio normalisation and non-AIDS-related events in individuals with HIV who achieve viral load suppression with antiretroviral therapy: an observational cohort study, The Lancet HIV 2 (3).

[20] V. Leung, J. Gillis, J. Raboud, C. Cooper, R. S. Hogg, M. R. Loutfy, et al., Predictors of CD4:CD8 ratio normalization and its effect on health outcomes in the era of combination antiretroviral therapy, PLoS One 8 (10).

[21] T. Le, E. J. Wright, D. M. Smith, W. He, G. Catano, J. F. Okulicz, et al., Enhanced CD4+ T-cell recovery with earlier HIV-1 antiretroviral therapy, New England Journal of Medicine 368 (3) (2013) 218–230.

[22] S. Serrano-Villar, T. Sainz, S. A. Lee, P. W. Hunt, E. Sinclair, B. L. Shacklett, et al., HIV-infected individuals with low CD4/CD8 ratio despite effective antiretroviral therapy exhibit altered T cell subsets, heightened CD8+ T cell activation, and increased risk of non-AIDS morbidity and mortality., PLoS Pathogens 10 (5) (2014).

[23] S. Serrano-Villar, S. Moreno, M. Fuentes-Ferrer, C. Sánchez-Marcos, M. Avila, T. Sainz, et al., The CD4:CD8 ratio is associated with markers of age-associated disease in virally suppressed HIV-infected patients with immunological recovery, HIV Medicine 15 (1) (2014) 40–49.

[24] S. Alizon, V. von Wyl, T. Stadler, R. Kouyos, S. Yerly, B. Hirschel, et al., Phylogenetic approach reveals that virus genotype largely determines HIV set-point viral load, PLoS Pathogens 6 (9).

[25] J. Fellay, D. Ge, K. Shianna, S. Colombo, B. Ledergerber, E. Cirulli, et al., Common genetic variation and the control of HIV-1 in humans, PLoS Genetics 5 (12) (2009).

[26] C. Fraser, K. Lythgoe, G. E. Leventhal, G. Shirreff, T. D. Hollingsworth, S. Alizon, et al., Virulence and pathogenesis of HIV-1 infection: An evolutionary perspective, Science 343 (6177).

[27] SHCS, Cohort Profile: The Swiss HIV Cohort Study, International Journal of Epidemiology 39 (5) (2010) 1179–1189.

[28] G. Greub, B. Ledergerber, M. Battegay, P. Grob, L. Perrin, H. Furrer, et al., Clinical progression, survival, and immune recovery during antiretroviral therapy in patients with HIV-1 and hepatitis C virus coinfection: the Swiss {HIV} Cohort Study, The Lancet 356 (9244) (2000) 1800 – 1805.

[29] N. M. Laird, J. H. Ware, Random-effects models for longitudinal data, Biometrics 38 (4) (1982) pp. 963–974.
URL http://www.jstor.org/stable/2529876

[30] P. Taffé, M. May, A joint back calculation model for the imputation of the date of HIV infection in a prevalent cohort, Statistics in Medicine 27 (23) (2008) 4835–4853.

[31] S. Nakagawa, H. Schielzeth, A general and simple method for obtaining R2 from generalized linear mixed-effects models, Methods in Ecology and Evolution 4 (2) (2013) 133–142.

[32] P. C. Johnson, Extension of Nakagawa & Schielzeth's R2GLMM to random slopes models, Methods in Ecology and Evolution 5 (9) (2014) 944–946.

[33] D. K. Pauler, The Schwarz criterion and related methods for normal linear models, Biometrika 85 (1) (1998) 13–27.

[34] J. Braun, L. Held, B. Ledergerber, the Swiss HIV Cohort study, Accounting for baseline differences and measurement error in the analysis of change over time, Statistics in Medicine 33 (1) (2014) 2–16.

[35] S. M. Hammer, K. E. Squires, M. D. Hughes, J. M. Grimes, L. M. Demeter, J. S. Currier, et al., A controlled trial of two nucleoside analogues plus Indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less, New England Journal of Medicine 337 (11) (1997) 725–733.

[36] R. D. Kouyos, V. von Wyl, S. Yerly, J. Böni, P. Taffé, C. Shah, et al., Molecular epidemiology reveals long-term changes in HIV type 1 subtype B transmission in Switzerland, Journal of Infectious Diseases 201 (10) (2010) 1488–1497.

[37] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, R Core Team, nlme: Linear and Nonlinear Mixed Effects Models, r package version 3.1-112 (2013).

[38] J. V. Giorgi, M. A. Majchrowicz, T. D. Johnson, P. Hultin, J. Matud, R. Detels, Immunologic effects of combined protease inhibitor and reverse transcriptase in-

hibitor therapy in previously treated chronic HIV-1 infection, AIDS 14 (12) (1998) 1833–44.

[39] D. C. Douek, L. J. Picker, R. A. Koup, T cell dynamics in HIV-1 infection, Annual Review of Immunology 21 (2003) 265–304.

[40] S. Muff, A. Riebler, L. Held, H. Rue, P. Saner, Bayesian analysis of measurement error models using integrated nested Laplace approximations, Journal of the Royal Statistical Society: Series C (Applied Statistics) 64 (2) (2015) 231–252.

[41] S. H. Panagiotakis, G. Soufla, S. Baritaki, G. Sourvinos, A. Passam, I. Zagoreos, et al., Concurrent CMV and EBV DNAemia is significantly correlated with a delay in the response to HAART in treatment-naive HIV type 1-positive patients, AIDS Research and Human Retroviruses 23 (1) (2007) 10 – 18.

[42] A. Ferraz da Silva, L. Giron, S. Ramos da Silva, A. Naime Barbosa, R. Almeida, D. Elgui de Oliveira, Human gammaherpesviruses viraemia in HIV infected patients, Journal of Clinical Pathology 68 (9) (2015) 726–732.

Supplementary material for "CD8 counts and CD4/CD8 ratio independently predict CD4 response in drug naive and in patients on cART"

Rafael Sauter, Ruizhu Huang, Bruno Ledergerber,
Huldrych F Günthard, Manuel Battegay, Enos Bernasconi,
Alexandra Calmy, Matthias Cavassini, Hansjakob Furrer,
Matthias Hoffmann, Leonhard Held and the Swiss HIV cohort study.

Email: rafael.sauter@uzh.ch

20th July 2015

1 Linear mixed model for longitudinal SHCS data

The SHCS data is collected for patients $i = 1, \ldots, m$ who are observed repeatedly at follow-up visits $j = 1, \ldots, n_i$ at times t_{i1}, t_{i2}, t_{ij}, where $j = 1, \ldots, n_i$. The square root transformed CD4 cell counts is the outcome $\sqrt{\text{CD4}_{ij}} = y_{ij}$ which is assumed to follow a normal distribution. The observations for a single patient i are modeled by a linear predictor

$$\mathbf{y}_i = \mathbf{x}_i\boldsymbol{\beta} + \mathbf{z}_i\mathbf{b}_i + \boldsymbol{\epsilon_i}.$$

The observations for this patient are collected by the vector $\mathbf{y}_i$ of length n_i. The p different fixed effect covariates are in the design matrix $\mathbf{x}_i$ of dimension $(n_i \times p)$. The coefficients for the fixed effect predictors are in the vector $\boldsymbol{\beta}$ and is of length p. The

patient-specific random effects vector $\mathbf{b}_i$ is usually a subvector of $\boldsymbol{\beta}$ with length $q < p$ and is multiplied by the random effects design matrix $\mathbf{z}_i$ of dimension $(n_i \times q)$.

The linear mixed model assumes that the residuals follow a normal distribution $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2_\epsilon)$ and the errors are independent and identical distributed. The patient-specific random effects $\mathbf{b}_i$ are also assumed to follow a normal distribution $\mathbf{b}_i \sim \mathcal{N}_q(\mathbf{0}, \mathbf{D})$. In the case of random intercept and random slope $q = 2$ and the random effects covariance matrix $\mathbf{D}$ is of dimension 2×2. Further the error term $\boldsymbol{\epsilon}_i$ and the random effects $\mathbf{b}_i$ are assumed to be independent. The patient-specific marginal likelihood integrates over the random effects $\mathbf{b}_i$ and follows a multivariate normal distribution with mean $\mathbf{x}_i\boldsymbol{\beta}$ and covariance matrix $\mathbf{z}_i\mathbf{D}\mathbf{z}_i^\top + \mathbf{I}_{n_i}$, where here $\mathbf{I}_{n_i}$ is the identity matrix of dimension n_i.

Model M1, described in the main text, included the following covariates for which a fixed effect coefficient was estimated:
$\mathbf{x}_{ij} = (1, t_{ij}, \sqrt{\text{CD4}_{i,(j-1)}}, \log(\text{RNA}_{i,(j-1)}), \text{AIDS}_{ij}, \text{age}_{ij}, \text{transmission}_i, \text{HCV}_i)$ and in the cART group additionally the effect for NRTI_i. Model M2 estimated the same fixed effects as M1 but additionally included $\sqrt{\text{CD8}_{i,(j-1)}}$. Model M3 estimated the same fixed effects as M1 but additionally included $\log(\text{CD4}_{i,(j-1)}/\text{CD8}_{i,(j-1)})$. The random effects structure is the same in all models and included a patient-specific random intercept and random slope such that $q = 2$ and the design matrix $\mathbf{z}_{ij} = (1, t_{ij})$.

1.1 Model choice criteria

Different models with different predictors can be compared by model choice criteria such as the AIC or BIC. Common model selection criteria, like AIC or BIC, are not applicable to linear mixed models. The main reason for this is the difficulty to assess the degree of freedom for the random effects included in a linear mixed model. Different proposals for extensions of the common model choice criteria to linear mixed models exist. One suggestion for a model choice criteria to mixed models with different fixed effects but the same random effects structure is the modified BIC [1], who defines the version of the BIC, modified for linear mixed models as

$$\text{BIC} = -2\log L + \sum_{k=1}^{p} \log(n_k)$$

where L is the likelihood and n_k is equal to the number of individuals I if the predictor is included as fixed and as random effect, or the number of observations $n = \sum_{i=1}^{I} n_i$ if the predictor is included as fixed effect only. See also [2] for a description of model choice criteria and about predictive comparisons for generalized linear mixed models.

1.2 Explained variation

A goodness of fit criteria for linear regression models is the R^2, which is a measure for the variation expressed by the model in relation to the overall variation in the data. In the case of a multiple regression model the R^2 needs to be adjusted for the number of included parameters, in order to gain comparability across different models with different parameters. For linear mixed models one needs to adapt the R^2 additionally for the variation explained by the random effects, which can be done in different ways.

Properties for a sensible R^2 measure are discussed by [3], who suggest a R^2 measure for linear mixed models with random intercepts. They distinguish between the marginal R^2, which expresses the variance explained by fixed effects as proportion of all variance components and the conditional R^2, which is a measure for the variance explained by fixed and random effects. This idea was extended to linear mixed models with random intercepts and random slopes by [4].

According to equation 29 in [3] the marginal R^2 for a random intercept model is defined as

$$R^2_{\text{marignal}} = \frac{\sigma_f^2}{\sigma_f^2 + \sum_{l=1}^{q} \sigma_l^2 + \sigma_\epsilon^2}$$

where σ_f^2 is the variance attributable to the fixed effects, σ_l^2 is the variance of the lth of the q random effects. For the extension to random intercepts and random slopes [4] proposes to replace $\sum_{l=1}^{q} \sigma_l^2$ by the mean random effect variance

$$\bar{\sigma}^2 = \text{tr}(\mathbf{z}\mathbf{D}\mathbf{z}^\top)/n$$

where n is the number of observations $(n = \sum_{i=1}^{I} n_i)$ and $\mathbf{z}$ the random effects design matrix for all patients of dimension $n \times 2$.

1.3 Time course for NAIVE and cART

In this section we illustrate in more detail how the applied linear mixed model for longitudinal data takes the time trend since cohort entry (NAIVE) or since therapy start (cART) into account. The time trend for both subgroups is illustrated in the upper two plots in Figure 1 which is based on the model M3 presented in the main text but would look rather similar for model M2. For Figure 1 all covariates are set to its average values and held constant and categorical variables are set to the reference category, while only the time since cohort entry for NAIVE or since therapy start for cART is increasing.

The fixed effect time trend, representing the population mean of the square root transformed CD4 course and given the reference categories for the transmission and the HCV factors, is plotted as black line. The dark grey area around the global fixed time effect shows a 95% confidence interval (CI) for the average population time trend. The light grey area shows a 95% CI for the time-dependent prediction error based on the standard deviation of the model residuals (σ_ϵ) and the standard deviation of the fixed time effect (σ_f). This prediction error band is based on the artificial data used for the plot, for which time since cohort entry (NAIVE) or since therapy start (cART) is the only varying covariate.

The linear decreasing time trend for NAIVE is opposed by a sharply increasing time trend in cART, as the time scale was square root transformed for the cART subgroup. This reflects that CD4 cells for untreated HIV infected patients are steadily declining where on average the CD4 level is recovering after cART initiation.

The dashed lines in the upper two plots of Figure 1 show patient-specific deviations from the population mean based on random intercepts and random slopes of 20 different, randomly sampled patients. These individual intercepts and time courses reflect how the linear mixed model takes the between patient variation at baseline and during disease progression into account. The model yields a rather high flexibility to cover individual time courses, especially just after cART initiation: a positive random slope on the square root transformed time scale leads to a sharp but flattening increase in CD4, especially if the CD4 level was already impaired which is reflected by a relatively small random intercept. A negative random slope instead is often present among the cART group if the CD4 level at therapy start was still high. A negative random slope implies a flattening decline in CD4 cell counts after cART initiation.

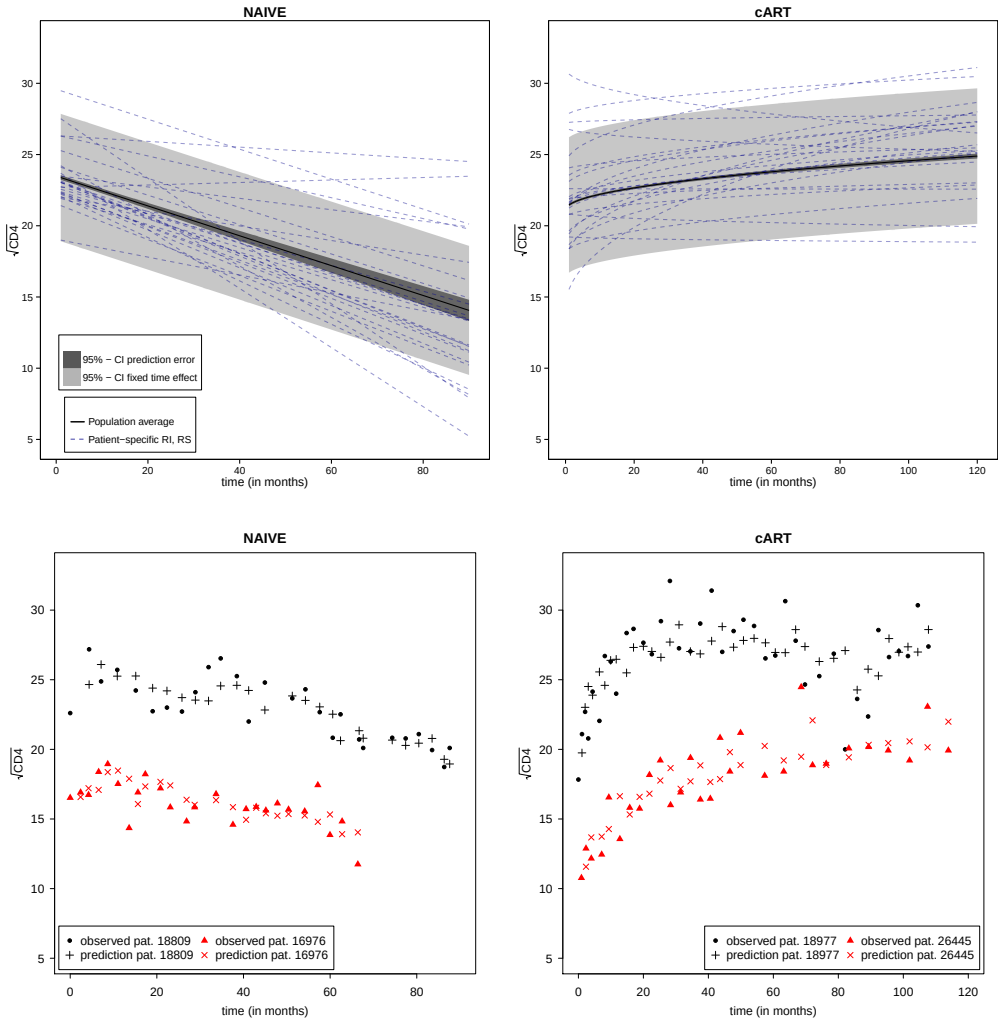

Figure 1: Profiles with time trend in the NAIVE and cART based on model M3 for the population average and 20 randomly selected patients (pat.) in the upper two plots. Observed values and model predictions for two selected patients in each group in the lower two plots.

The lower two plots in Figure 1 show observed square root transformed CD4 cell counts obtained from two NAIVE and two cART patients, as well as each correspond-

ing model prediction for disease progression. The four patients in the lower plots are chosen such that their response values do not overlap and that they have a substantial record of observed lymphocytes. The lower two plots serve for illustration purposes only and should clarify how the model applies to the patient-specific longitudinal data.

1.4 Random effect structure and residual analysis

The parameter estimates of the random effect covariance matrix ($\hat{\mathbf{D}}$) are shown in Table 1 based on model M2 and model M3, presented in the main text, together with the residual standard deviations ($\hat{\sigma}_{\boldsymbol{\epsilon}}$). The estimated random effect structure and the standard deviation for the residuals are very similar for both models. From Table 1 we see that the patient-specific variation of the intercept has about the same magnitude as the residual standard deviation. The size of the residual standard deviation for model M3 is also visualized in Figure 1 in the 95% prediction error band as light grey area.

	NAIVE		**cART**	
	M2	**M3**	**M2**	**M3**
RI Stdev.	2.272	2.267	3.330	3.130
RS Stdev.	0.093	0.090	0.517	0.502
correlation	-0.260	-0.309	-0.685	-0.646
Resid. Stdev.	2.281	2.280	2.430	2.425

Table 1: Estimated random effect structure and residual standard error of linear mixed models M2 and M3 for each patient subgroup (NAIVE and cART).

Figure 2 shows a QQ-plot of the residuals for models M2 and M3 and the NAIVE and cART patient groups. The residuals in Figure 2 and 3 are raw residuals divided by the corresponding standard errors and further normalized by the inverse square-root factor of the estimated error correlation matrix (see `residuals.lme` in the R-package `nlme` for more information).

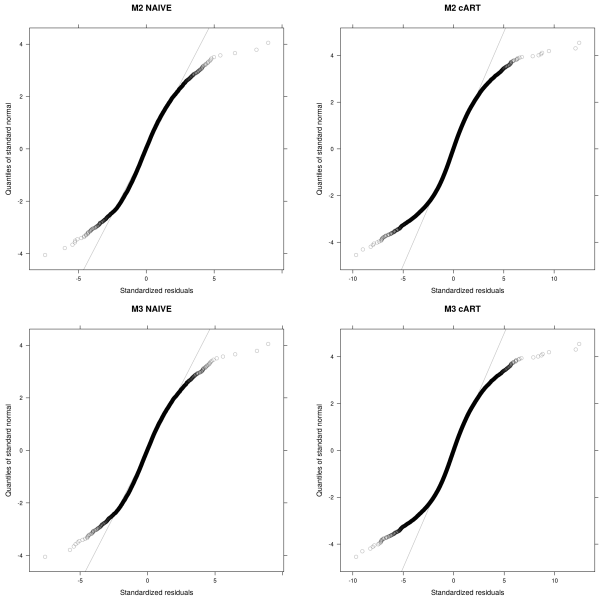

Figure 2: QQ-plots with standardized reisduals in model M2, M3 for NAIVE, cART.

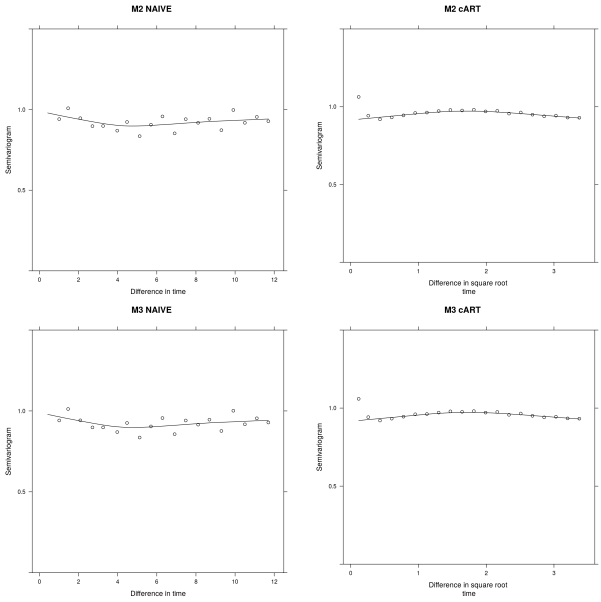

Figure 3: Variogram with standardized reisduals in model M2, M3 for NAIVE, cART.

2 Effect size of predictors

To overcome the difficulties in giving an interpretation of the model coefficients, which are transformed to different non-linear scales, we illustrate the effect sizes of the predictors and the transformation from $\sqrt{\text{CD4}}$ to original CD4 cell counts with a plots in Figure 4 and 5 for model M2 and in Figure 6 and 7 for model M3. The horizontal bars indicate the lower 2.5% and the upper 97.5% quantile as well as the median (black dot) of each covariate multiplied by its coefficient estimate, labelled with the corresponding quantiles of the covariates. As the level effect sizes for the categorical predictors HCV and transmission are very small, they are omitted in Figure 4 to 7.

The contribution of each predictor to the square root transformed response (here $\sqrt{\text{CD4}}$) is additive and can be read off for every predictor from the line at the bottom (predicted $\sqrt{\text{CD4}}$). Summing up each contribution to the response for all predictors gives the prediction of the square root transformed CD4 cell counts at the next follow-up visit. The vertical scale on the right allows to translate from the predicted square root scale (small figures) to the CD4 cell counts (large figures). On the right vertical scale also the fixed effect intercept, which must be added to the predicted value, is indicated by a black dot. The error bar for the fixed effect intercept is representing a 95% CI based on the estimated random intercept standard deviation. One can read off from the scale for the absolute CD4 cell counts on the right, that the same difference in the linear predictor causes a larger difference in the absolute CD4 cell count, if the predicted square root CD4 cell count is higher.

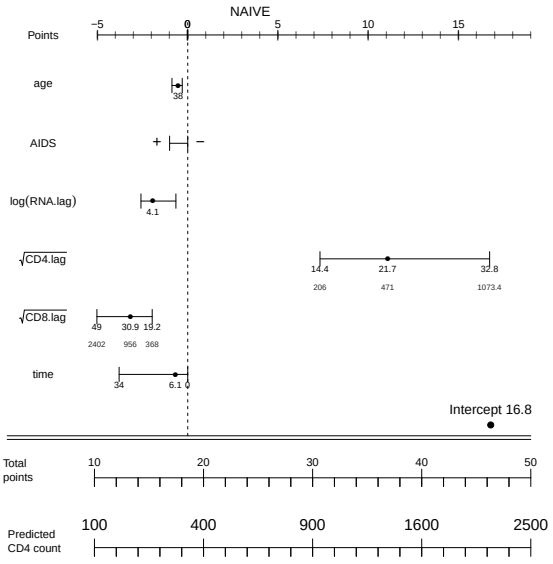

Figure 4: Effect size of predictors for NAIVE patients and model M2.

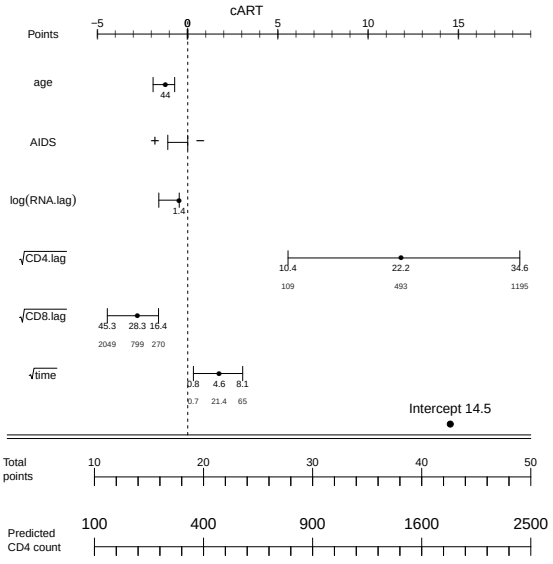

Figure 5: Effect size of predictors for cART patients and model M2.

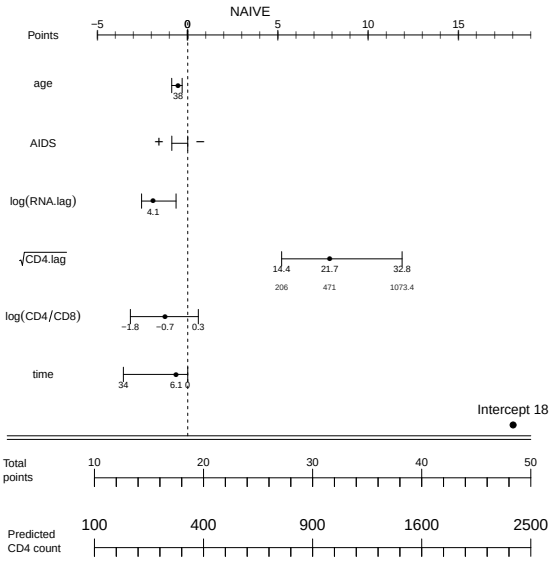

Figure 6: Effect size of predictors for NAIVE patients and model M3.

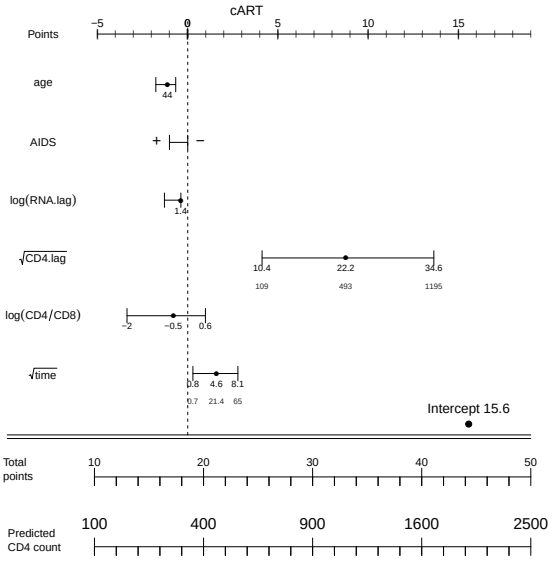

Figure 7: Effect size of predictors for cART patients and model M3.

3 Coefficient estimates for Model M2

Table 2 shows the coefficient estimates for the longitudinal predictors for model M2, corresponding to Table 3 in the main text, which reported the estimates for model M3. Equivalently Table 3 shows the coefficient estimates for the additional coefficients based on Model M2, corresponding to Table 4 in the main text.

	NAIVE		**cART**	
	coef.	95% - CI	coef.	95% - CI
intercept	16.78	16.21 to 17.36	14.54	14.30 to 14.78
time	-0.11	-0.12 to -0.10		
square root time			0.38	0.36 to 0.40
CD4	0.51	0.50 to 0.52	0.53	0.53 to 0.54
RNA	-0.47	-0.53 to -0.40	-0.34	-0.36 to -0.32
CD8	-0.10	-0.11 to -0.09	-0.10	-0.10 to -0.09

Table 2: Coefficient estimates (coef.) with 95% confidence intervals (CI) for longitudinal predictors and both patient subgroups (NAIVE and cART) for model M2. All p-values are <0.0001. Estimates for additional predictors are in Table 3.

	NAIVE			**cART**		
	coef.	95% CI	p-value	coef.	95% CI	p-value
AIDS at follow-up visit	-1.00	-1.41 to -0.60	< 0.0001	-1.10	-1.21 to -0.99	< 0.0001
age at follow-up visit	-0.01	-0.02 to 0.00	0.0023	-0.03	-0.03 to -0.02	< 0.0001
NRTI at baseline				-0.02	-0.02 to -0.01	< 0.0001
transmission			0.01			< 0.0001
MSM (reference)	0.00			0.00		
IDU-male	-0.13	-0.60 to 0.34		-1.19	-1.45 to -0.93	
IDU-female	-0.11	-0.70 to 0.49		-0.73	-1.04 to -0.42	
HET-male	-0.35	-0.63 to -0.07		-0.64	-0.79 to -0.48	
HET-female	-0.40	-0.65 to -0.14		-0.44	-0.58 to -0.29	
other	-0.07	-0.58 to 0.43		-0.34	-0.61 to -0.08	
HCV			0.39			< 0.0001
negative (reference)	0.00			0.00		
inactive	-0.09	-0.61 to 0.42		0.02	-0.28 to 0.33	
active	-0.16	-0.52 to 0.21		-0.45	-0.66 to -0.24	

Table 3: Estimates for additional predictors for model M2.

4 Time spans between follow-up visits

Figure 8 shows a histogram for the time between two subsequent follow-up visits for the NAIVE and the cART subgroups. The vertical dashed line indicates a follow-up time of 12 months. All observations with a follow-up time of more than one year were censored.

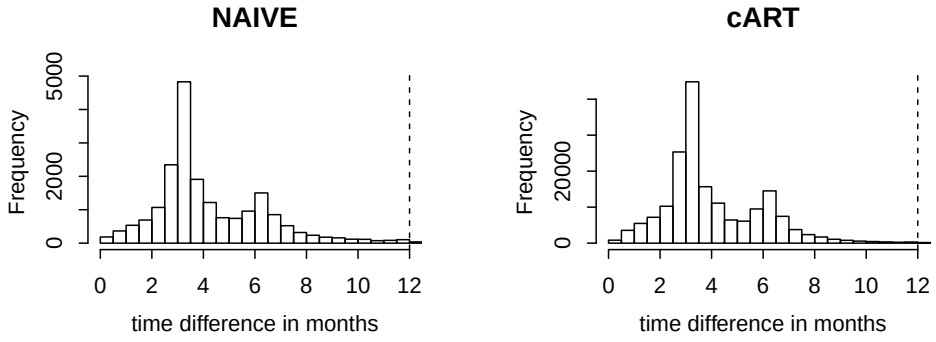

Figure 8: Histograms for time differences between two subsequent follow-up visits in months. Time is equal to zero at cohort entry for NAIVE and at therapy initiation for cART.

References

[1] D. K. Pauler, The Schwarz criterion and related methods for normal linear models, Biometrika 85 (1) (1998) 13–27.
URL http://biomet.oxfordjournals.org/content/85/1/13.abstract

[2] J. Braun, L. Held, B. Ledergerber, Predictive Cross-validation for the Choice of Linear Mixed-Effects Models with Application to Data from the Swiss HIV Cohort Study, Biometrics 68 (1) (2012) 53–61. doi:10.1111/j.1541-0420.2011.01621.x.
URL http://dx.doi.org/10.1111/j.1541-0420.2011.01621.x

[3] S. Nakagawa, H. Schielzeth, A general and simple method for obtaining R2 from

generalized linear mixed-effects models, Methods in Ecology and Evolution 4 (2) (2013) 133–142. doi:10.1111/j.2041-210x.2012.00261.x.
URL http://dx.doi.org/10.1111/j.2041-210x.2012.00261.x

[4] P. C. Johnson, Extension of Nakagawa & Schielzeth's R2GLMM to random slopes models, Methods in Ecology and Evolution 5 (9) (2014) 944–946. doi:10.1111/2041-210X.12225.
URL http://dx.doi.org/10.1111/2041-210X.12225

PAPER II

Quasi-complete Separation in Random Effects of Binary Response Mixed Models: Integrated Nested Laplace Approximations vs. MCMC

Rafael Sauter, Leonhard Held

Paper published in *Journal of Statistical Computation and Simulation*.

Quasi-complete Separation in Random Effects of Binary Response Mixed Models

R. Sauter[a*] and L. Held[a]

[a] *Department of Biostatistics, Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Hirschengraben 84, 8001 Zürich*

Clustered observations such as longitudinal data are often analysed with generalized linear mixed models (GLMM). Approximate Bayesian inference for GLMMs with normally distributed random effects can be done using integrated nested Laplace approximations (INLA), which is in general known to yield accurate results. However, INLA is known to be less accurate for GLMMs with binary response. For longitudinal binary response data it is common that patients do not change their health state during the study period. In this case the grouping covariate perfectly predicts a subset of the response, which implies a monotone likelihood with diverging maximum likelihood (ML) estimates for cluster-specific parameters. This is known as quasi-complete separation. In this paper we demonstrate, based on longitudinal data from a randomized clinical trial and two simulations, that the accuracy of INLA decreases with increasing degree of cluster-specific quasi-complete separation. Comparing parameter estimates by INLA, Markov chain Monte Carlo sampling and ML shows that INLA increasingly deviates from the other methods in such a scenario.

Keywords: integrated nested Laplace approximations; Bayesian generalized mixed models; cluster-specific quasi-complete separation

1. Introduction

There has been recent interest in the accuracy of integrated nested Laplace approximations (INLA) [1] for Bayesian inference in binary response mixed models. INLA has been successfully applied to generalized linear mixed models (GLMMs) [2], and a generally high accuracy has been reported. However, for the special case of binary responses, a thorough comparison with Markov chain Monte Carlo (MCMC) sampling has identified larger discrepancies [2]. Here, the relative approximation error, measured as the difference between the marginal posterior mean with MCMC and INLA, and scaled with the (MCMC) posterior standard deviation, was around 30%. These results are in contrast to a more recently published simulation study [3], which reported a high accuracy of INLA.

There is also interest in the accuracy of classical maximum likelihood (ML) estimates in GLMMs with binary responses. ML inference requires numerical integration over the random effects, for which penalized quasi likelihood (PQL) or adaptive Gauss Hermite quadrature (GHQ) are the two most common approaches. In response to the increasing usage of GLMMs in ecology and evolution, an overview of commonly used software packages for GLMMs has been published [4]. A detailed comparison of the estimates obtained by different statistical software packages has identified substantial differences [5], *e. g.* between `PROC NLMIXED` in `SAS` and the function `glmer()` in `R`, although both use adaptive GHQ integration. Also, the accuracy of Bayesian and ML estimation methods has been compared [6], who also consider results with INLA produced in the simulation study by [2].

*Corresponding author. Email: rafael.sauter@uzh.ch

Unfortunately, there is no analytical expression for the approximation error of INLA [1]. A straightforward way to assess INLA's accuracy is a direct comparison with MCMC. Alternatively, the accuracy of INLA in binary response models has been contrasted with the computationally more intensive expectation propagation (EP) algorithm [1] originating from the machine learning literature [7, 8].

There seems to be room for further comparisons of INLA and MCMC in other scenarios than investigated so far. We challenge INLA with a special but still realistic situation, in which not only INLA but also other estimation methods may run into problems. Specifically, we consider a situation, where a covariate is (almost) perfectly classifying the response, known as (quasi) complete separation [9]. For longitudinal data with binary response, cluster-specific (quasi) complete separation may occur if a patient shows no variation in the response, *i. e.* has longitudinal profile $(0, \ldots, 0)$ or $(1, \ldots, 1)$. Based on longitudinal data from a clinical trial on the presence of toenail infections and an additional simulation study, we show that in this case the INLA parameter estimates do not agree with those obtained by MCMC or ML. Further we assess the root mean squared error and bias of the parameter estimates by INLA in a second simulation study. Cluster-specific quasi-complete separation causes a bias for INLA which implies a substantially lower accuracy than reported elsewhere [2, 3, 6].

This paper is organized as follows. We start by reviewing likelihood and Bayesian inference in GLMMs in Section 2. In Section 3, we empirically compare parameter estimates obtained from applying INLA, MCMC and ML to the toenail clinical trial data. Section 4 describes results from two simulation studies with varying degree of cluster-specific quasi-complete separation. We close with some discussion in Section 5.

2. Inference for binary response mixed model

Consider a GLMM for (possibly unbalanced) longitudinal data with binary response $y_{ij} \in \{0, 1\}$ from individuals $i = 1, \ldots, I$ at occasions $j = 1, \ldots, n_i$, linked to times t_{ij} at which the measurements are taken. The total number of observations is $n = \sum_i^I n_i$. The logistic mixed model

$$\text{logit}(\pi_{ij}) = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{z}_{ij}^\top \mathbf{b}_i$$

assumes that the binary observations y_{ij} are conditionally independent, given the random effects $\mathbf{b}_i$, with success probability $\pi_{ij} = \Pr(y_{ij} = 1 \,|\, \boldsymbol{\beta}, \mathbf{b}_i, \mathbf{D})$. Here $\mathbf{x}_{ij}$ is a vector of length p with explanatory variables and associated fixed effects vector $\boldsymbol{\beta}$. The cluster-specific random effects $\mathbf{b}_i$ are linked to the covariate vector $\mathbf{z}_{ij}$ of length q. The random effects are assumed to follow a multivariate normal distribution, *i. e.* $\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$. In the random intercept (RI) model, $q = 1$, $z_{ij} = 1$ and $\mathbf{D}$ is defined by only one hyperparameter $\delta = \sigma_b^2$, the variance of the random intercept. For a random intercept and slope model (RI+RS), $q = 2$, $z_{ij} = (1, t_{ij})^\top$ and the covariance matrix $\mathbf{D}$ consists of three hyperparameters $\boldsymbol{\delta}$, two random effect variances on the diagonal and the corresponding correlation.

2.1. *Likelihood inference*

Likelihood inference is based on the marginal likelihood of the GLMM. The marginal likelihood contribution of individual i is

$$f(\boldsymbol{y}_i \,|\, \boldsymbol{\beta}, \mathbf{D}) = \int \prod_{j=1}^{n_i} f(y_{ij} \,|\, \boldsymbol{\beta}, \mathbf{b}_i) f(\mathbf{b}_i \,|\, \mathbf{D}) d\mathbf{b}_i \tag{1}$$

where $f(\cdot)$ denotes either a probability mass or a density function and $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^\top$ is the response vector of the i-th individual.

Corresponding to $\mathbf{y}_i$ the matrices $\mathbf{x}_i$ and $\mathbf{z}_i$ collect the fixed and random effect vectors for all n_i observations and are of dimension $n_i \times p$ and $n_i \times q$. In a linear mixed model, the individual marginal likelihood follows a multivariate normal distribution with mean equal to $\mathbf{x}_i\boldsymbol{\beta}$ and covariance matrix $\mathbf{z}_i\mathbf{D}\mathbf{z}_i^\top + \sigma^2\mathbf{I}_{n_i}$, here $\mathbf{I}_{n_i}$ is the identity matrix of dimension n_i. This is not the case for a GLMM with non-normal response, where numerical integration over the q-dimensional vector $\mathbf{b}_i$ is required to compute (1). This task is usually solved by numerical integration *e. g.* via the Laplace approximation [10]. An alternative approach is based on PQL [11], where bias-corrections are available [12, 13], or the GHQ-approximation, which can be improved by selecting the points, at which the function is evaluated, adaptively [14]. Increasing the number of quadrature points also increases the accuracy of this approximation. With a single quadrature point the GHQ-approximation reduces to the Laplace approximation. In practice, numerical optimization of the marginal likelihood with respect to $\boldsymbol{\beta}$ and $\mathbf{D}$ is performed with random effects fixed at the empirical Bayes estimates $\tilde{\mathbf{b}}_i$. Finding $\tilde{\mathbf{b}}_i$ for fixed $\boldsymbol{\beta}$ and $\mathbf{D}$ is the first step and numerical optimization of the approximated likelihood is the second step, which both are iteratively updated until convergence is reached.

2.2. *Bayesian inference*

A Bayesian GLMM is a hierarchical model with three stages. The first stage is a model $f(\mathbf{y} \,|\, \boldsymbol{\theta})$ for the observed data $\mathbf{y}$, given the unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \mathbf{b}_1^\top, \ldots, \mathbf{b}_I^\top)^\top$. The second stage $f(\boldsymbol{\theta} \,|\, \boldsymbol{\delta})$ is the distribution of $\boldsymbol{\theta}$, given unknown hyperparameters $\boldsymbol{\delta}$. For a GLMM the distribution $f(\boldsymbol{\theta} \,|\, \boldsymbol{\delta})$ is assumed to be Gaussian, such that the GLMM can be described as a Gaussian Markov random field (GMRF) with precision matrix $\mathbf{Q}(\boldsymbol{\delta}) = \mathbf{D}(\boldsymbol{\delta})^{-1}$ [15]. The GMRF is controlled by a relatively small number of hyperparameters $\boldsymbol{\delta}$. The corresponding prior distribution $f(\boldsymbol{\delta})$ is the third stage of the formulation. In GLMMs, the hyperparameters $\boldsymbol{\delta}$ describe the covariance structure of the random effects. The posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ is

$$\begin{aligned} f(\boldsymbol{\theta}, \boldsymbol{\delta} \,|\, \mathbf{y}) &\propto f(\boldsymbol{\delta}) f(\boldsymbol{\theta} \,|\, \boldsymbol{\delta}) \prod_{i=1}^{I} f(\mathbf{y}_i \,|\, \boldsymbol{\theta}, \boldsymbol{\delta}) \\ &\propto f(\boldsymbol{\delta}) \,|\, \mathbf{Q}(\boldsymbol{\delta}) \,|^{\frac{1}{2}} \exp\left\{ -\frac{1}{2}\boldsymbol{\theta}^\top \mathbf{Q}(\boldsymbol{\delta})\boldsymbol{\theta} + \sum_{i=1}^{n} \log f(\mathbf{y}_i \,|\, \boldsymbol{\theta}, \boldsymbol{\delta}) \right\} \end{aligned}$$

and one of the major goals is to calculate the marginal posterior distribution of the k^{th} component of $\boldsymbol{\theta}$:

$$f(\theta_k \,|\, \mathbf{y}) = \int_{\boldsymbol{\delta}} \int_{\boldsymbol{\theta}_{-k}} f(\boldsymbol{\theta}, \boldsymbol{\delta} \,|\, \mathbf{y}) d\boldsymbol{\theta}_{-k} d\boldsymbol{\delta} = \int_{\boldsymbol{\delta}} f(\theta_k \,|\, \boldsymbol{\delta}, \mathbf{y}) f(\boldsymbol{\delta} \,|\, \mathbf{y}) d\boldsymbol{\delta}, \tag{2}$$

here $\boldsymbol{\theta}_{-k}$ denotes all components of $\boldsymbol{\theta}$ except the k^{th} one. Usually MCMC sampling is used to generate samples from $f(\theta_k \,|\, \mathbf{y})$. A binary response GLMM may require advanced sampling algorithms such as block updating [16, 17]. The computationally less intensive INLA approach [1] approximates the marginal posterior distributions by first applying a Laplace approximation [10] to the posterior distribution of $\boldsymbol{\delta}$ and a second Laplace approximation to the posterior of the components of $\boldsymbol{\theta}$ for selected values of $\boldsymbol{\delta}$. INLA uses numerical integration over the hyperparameters to finally obtain the marginal posterior distributions $f(\theta_k \,|\, \mathbf{y})$ of all components of $\boldsymbol{\theta}$. Three different approximation strategies to the first component $f(\theta_k \,|\, \boldsymbol{\delta}, \mathbf{y})$ in equation (2) are discussed in [1]: the first is the least accurate and uses a Gaussian approximation, the second is more precise and computationally more intensive and applies a Laplace approximation while the third is intermediate in accuracy and computing time and uses a simplified Laplace approximation. For all computations involving INLA, we used the intermediate simplified Laplace approximation strategy.

Bayesian inference requires specification of a prior distribution for $f(\boldsymbol{\beta})$ and $f(\boldsymbol{\delta})$. A common approach, also employed in this paper, are independent normal distributions with large variance, *e. g.* $1/\sigma_\beta^2 = 0.0001$, for each component of $\boldsymbol{\beta}$. In the RI model, we follow the approach by [2] and use an inverse gamma $\text{IG}(a_1, a_2)$ prior [18] for the variance σ_b^2. Integration over the hyperparameter for a normal distributed $f(b_i \,|\, \sigma_b^2)$ results in a marginal $t(0, a_2/a_1, 2a_1)$ distribution [18]. For this marginal t distribution a range is defined, which covers the odds ratio $\exp(b_i)$ with a probability of 95%. The values $a_1 = 0.5$ and $a_2 = 0.00802$ for the inverse Gamma prior $f(\sigma_b^2)$ are derived from the relationship between the marginal t distribution $f(b_i)$ and the assumed range for $\exp(b_i)$, which is $[0.2, 5]$ in this case. The same derivation with a range of $[0.1, 10]$ for $\exp(b_i)$ was used by [2, 3]. As discussed by [2] the same approach to determine an informative prior can be extended to the RI+RS model. In the RI+RS model, the covariance matrix $\mathbf{D}$ is assumed to follow an inverse Wishart $\text{IWi}(r, \mathbf{R})$ distribution [18], where $r = 5$ and $\mathbf{R}$ is a diagonal matrix with entries equal to 1.34.

2.3. *Quasi-complete separation*

Fitting a logistic regression model is problematic if a covariate perfectly predicts the response. Such a covariate implies that the ML estimate will be infinite as the likelihood is increasing monotonically. Although a perfect predictor is desirable, one would rarely accept such an extreme estimate based on a finite sample. The problem of divergent ML estimates for such a data configuration is defined as complete separation [9]. A weaker form is quasi-complete separation which occurs if the covariate predicts a subset of the response vector perfectly. Quasi-complete separation leads to infinite ML estimates for the covariate almost perfectly predicting the response but not for additional covariates, if present, which explain the remaining variation in the response.

In the particular case of a binary covariate, which completely separates the response, the corresponding 2×2 table has no off-diagonal entries. For a quasi-complete separation only one of the off-diagonal entries would be zero. A continuous covariate implies complete separation if *e. g.* for all negative values the response is one and for all positive values the response is zero. Quasi-complete separation is present if additionally for covariate values equal to zero the response is either one or zero.

Divergent ML estimates caused by complete separation in generalized linear models (GLMs) may be addressed by a penalized likelihood approach [19]. The suggested penalization depends on the inverse Fisher information matrix and is related to Jeffreys' invariant prior [19]. For a logistic regression with a completely separating binary covariate this approach corresponds to adding 1/2 to each cell of the 2×2 table. While

removing the small sample bias the penalized likelihood approach yields consistent estimates [19] and there exist different approaches to improve the coverage probability of the corresponding confidence intervals [20, 21].

Addressing quasi-complete separation in a logistic regression model with random intercept is discussed by [22]. However, complete separation may now be not only fixed covariate-specific but also be cluster-specific, affecting the random effects $\mathbf{b}_i$. More specifically, if the grouping covariate, which defines the random effect clusters, is separating the response, we encounter a cluster-specific complete separation for the random intercepts. For a logistic mixed model this occurs if all components of $\mathbf{y}_i$ are either equal to one or equal to zero. We have a cluster-specific quasi-complete separation if this occurs only for some i but not all I clusters.

The assumption $\mathbf{b}_i \sim \mathrm{N}(\mathbf{0}, \mathbf{D})$ penalizes deviations of $\mathbf{b}_i$ away from zero which in the case of cluster-specific quasi-complete separation stabilizes the marginal likelihood such that the estimates for $\mathbf{b}_i$ are defined. But the penalization decreases if the covariance matrix of the random effects $\mathbf{D}$ increases such that the parameter estimates $\mathbf{b}_i$ are not treated different from the fixed effects if $\mathbf{D}^{-1} \to \mathbf{0}$. Thus in the extreme case of cluster-specific complete separation, the ML estimates for $\mathbf{b}_i$ will not be defined, as the penalization term vanishes with the random effects variance going to infinity. For a random intercept plus random slope model the penalization term may be increased through the random effects correlation, if only one of the two random effect covariates is causing quasi-complete separation.

Depending on the degree of cluster-specific quasi-complete separation, *i. e.* the proportion of clusters with constant response, convergence problems will arise in the numerical optimization algorithms described in Section 2.1. Also depending on how many clusters are perfectly predicted by the grouping covariate, the normal assumption for the random effects may be inappropriate. Indeed, random effect estimates tend to have extreme values in the presence of cluster-specific quasi-complete separation.

Bayesian inference for GLMMs addresses the complete separation problem in random effects by an additional, possibly informative prior $f(\boldsymbol{\delta})$ [23]. The prior distribution $f(\boldsymbol{\delta})$ needs to be proper [24], so the posterior $f(\boldsymbol{\theta}, \boldsymbol{\delta} \,|\, \mathbf{y})$ will also be proper. Nevertheless, even for Bayesian inference, numerical problems may arise with increasing degree of quasi-complete separation.

3. INLA vs. MCMC for toenail infection data

The data considered in this section are the result from a randomized, double-blinded clinical trial comparing two oral treatments for toenail infections [25–27]. The data are available on `http://onlinelibrary.wiley.com/journal/10.1111/%28ISSN%291467-9876/homepage/50_3.htm`. The primary response was the degree of onycholysis, *i. e.* the degree of separation of the nail plate from the nail-bed. The response was classified into absent, mild, moderate or severe onycholysis and was further aggregated to a binary response with either absent or mild (0, not severe) or moderate to severe (1, severe) degree [26, 27].

Follow-up visits were planned to take place 1, 2, 3, 6, 9 and 12 months after baseline. However, the actual times t_{ij} of follow-up visits varied around the foreseen schedule and some patients have less than 6 follow-up measurements due to drop out. For the following analysis 5 patients with no follow-up measurements have been removed such that the dataset consists of 1903 observations from 289 individuals. There are 160 patients who stay always in the not severe state throughout the observation period and 14 patients who remain always in the severe state, while all remaining 115 patients change their

disease state at least once. Time since baseline was centred at the overall mean in order to improve the mixing of the MCMC algorithm [23]. The fixed effects for all models consist of an intercept, the treatment effect, the centred time since baseline in months and the interaction for centred time and treatment, *i. e.* $\mathbf{x}_{ij} = (1, \text{trt}_i, t_{ij}, t_{ij} \times \text{trt}_i)^\top$. The toenail infection data is analysed with a binary response RI ($z_{ij} = 1$) and with a RI+RS ($\mathbf{z}_{ij} = (1, t_{ij})^\top$) model. The RI model has only one hyperparameter which is the random intercept variance σ_b^2. For the RI+RS model the random effect covariance matrix $\mathbf{D}$ is defined by three hyperparameters: the variance for the random intercept $\sigma_{b_1}^2$, the variance for the random slope $\sigma_{b_2}^2$ and the correlation parameter between the two variances ρ.

INLA is implemented in a software package and an `R`-interface is available on `http://www.r-inla.org/`. We used the `r-inla` version built on 14. July 2014. All MCMC sampling was done with `JAGS` [28] through the `R`-interface `R2jags` and the `R`-package `coda` [29]. For binary or binomial response data, `JAGS` uses the algorithms proposed by [17] and [30]. Still we used a relatively large number of 500'000 MCMC iterations with 20'000 additional burnin iterations and thinning of 200 in both models, the RI and RI+RS model, to reach convergence and to ensure a negligible Monte Carlo error of the parameter estimates. ML estimation of the models was undertaken with the `R` package `lme4` [31], version 0.999999-2. We did not use the latest `lme4` version because it restricts the maximal number of quadrature points in the GHQ-approximation to 25 for the RI model and to 1 for the RI+RS model. In the RI model we use 20 quadrature points for the one-dimensional integration and the results are the same as with the current `lme4` version, whereas we use 50 quadrature points for the two-dimensional integration over the joint random effects distribution in the RI+RS model. All computations were done with `R` version 2.15.3 (2013-03-01). See Appendix A for more details about the influence of the number of quadrature points for the toenail data models.

We compare the ML estimates with the marginal posterior means for all components of $\boldsymbol{\beta}$, while fixing the hyperparameters $\boldsymbol{\delta}$ for the Bayesian methods at the ML estimates. Under an uninformative prior for the components of $\boldsymbol{\beta}$ and without any uncertainty in the hyperparameters, the posterior means should be very close to the ML estimates. The only difference between the ML estimate and the mean of the marginal posterior distribution of a fixed effect is the integration over both, random and the remaining fixed effects, while the marginal likelihood only integrates over random effects. Alternatively, the (joint) posterior mode could be used, but this is not the standard output for `r-inla`, which is approximating the marginal posteriors. Anyhow, posterior means and modes will coincide to a reasonable accuracy, since the posterior of $\boldsymbol{\beta}$ is known to be asymptotically Gaussian.

3.1. *Differences in the parameter estimates*

The estimated marginal posterior densities of $\boldsymbol{\beta}$ for both the RI and the RI+RS model are shown in Figure 1. Both are obtained using a fully Bayesian approach with hyperpriors for $\boldsymbol{\delta}$ as described in Section 2.2. Each histogram is based on the MCMC samples provided by `JAGS` and the lines show the corresponding marginal posterior density estimate produced by `r-inla`. For the RI model in the upper row of Figure 1 we see that MCMC and INLA agree rather well for all fixed effects, except for the time covariate, where there is a slight shift towards zero for the posterior by INLA compared to the MCMC histogram. In the lower row of Figure 1 we see more substantial differences between INLA and MCMC for the treatment effect and the interaction between time and treatment.

Figure 2 shows the posterior distributions of the same fixed effects and the same models as in Figure 1 but now with hyperparameters fixed at values which were determined

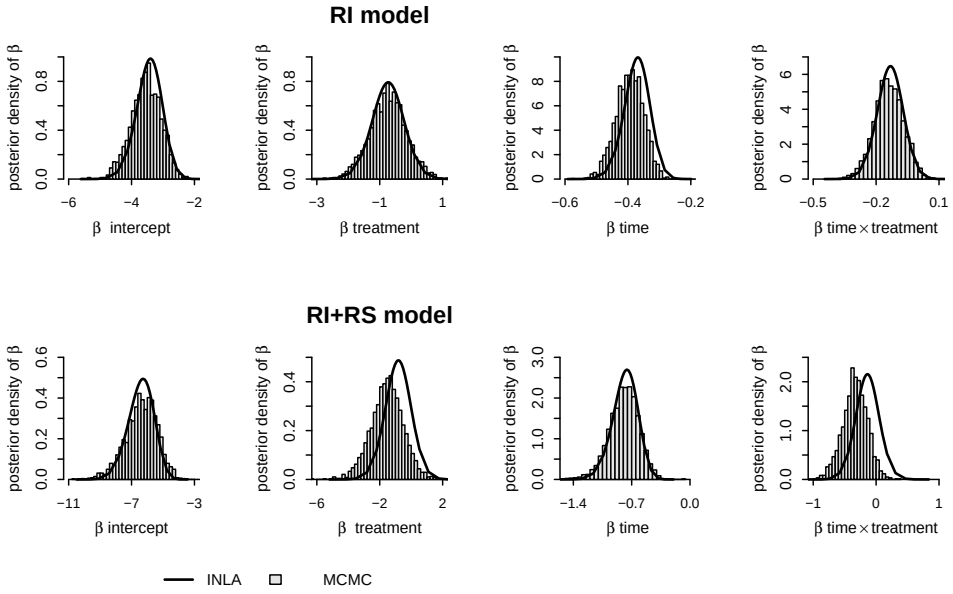

Figure 1. Marginal posterior distributions of fixed effects $\boldsymbol{\beta}$ with MCMC (histogram) and INLA (line).

by ML with `lme4`. The additional red lines now give approximate normal "posterior" distributions based on the ML estimates and the corresponding standard errors. In all plots of Figure 2 we see that the approximate posterior distributions based on the ML es-

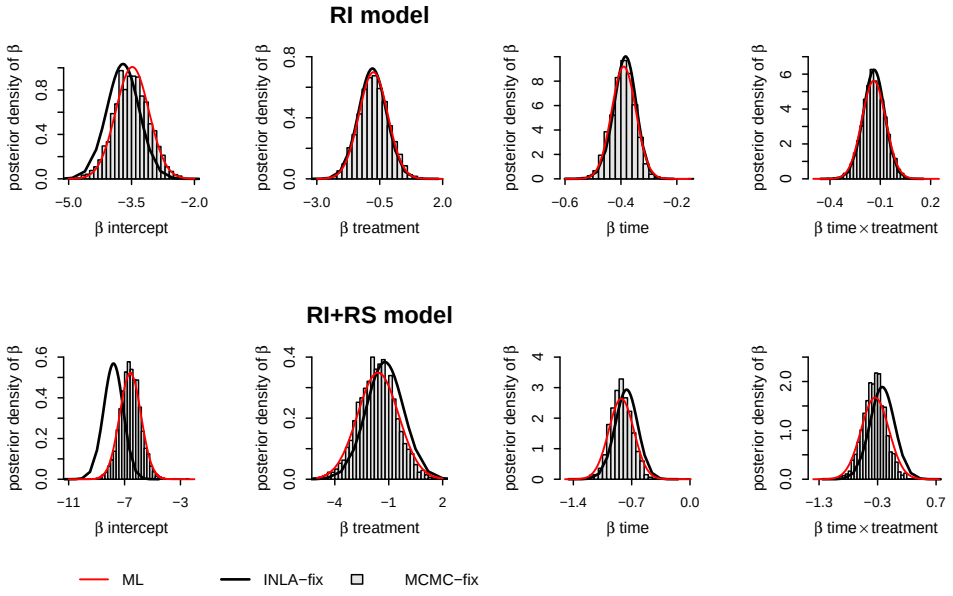

Figure 2. Marginal posterior distributions of $\boldsymbol{\beta}$ by MCMC (histogram), INLA (black line) and based on ML estimates (red line). Hyperparameters values are fixed at corresponding ML estimates.

timates agree well with the MCMC histograms. However, the posterior density estimates provided by INLA exhibit a substantial bias for the intercepts of both the RI and the RI+RS model. There is also some discrepancy for the other fixed effects in the RI+RS model.

The upper half of Table 1 summarizes the differences between the posterior mean estimates obtained with INLA or MCMC and with ML estimates. In the lower half of Table 1, relative differences are given, by scaling the differences from the upper part with the MCMC marginal posterior standard deviation, in the same way as done in the simulation study by [2]. The left part of Table 1 reports differences for the RI, the right half for the RI+RS model. For comparison with the ML estimates, we fixed the hyperparameter of the RI model at the ML estimate σ_b^2=16.04. For the RI+RS model the random intercept variance was fixed at $\sigma_{b_1}^2$ =47.75, the random slope variance at $\sigma_{b_2}^2 = 1.04$ and the correlation at ρ = -0.05.

Table 1. Differences (top) and relative differences (bottom) between parameter estimates obtained with MCMC, INLA and ML for the RI (left) and the RI+RS (right) model. Relative differences are scaled with the MCMC marginal posterior standard deviation. Comparisons of INLA and MCMC with ML are based on hyperparameter values fixed at the corresponding ML estimates.

	RI model			RI+RS model		
	MCMC	ML	ML	MCMC	ML	ML
	INLA	INLA-fix	MCMC-fix	INLA	INLA-fix	MCMC-fix
intercept	-0.073	0.225	0.014	0.101	1.228	0.038
treatment	-0.041	0.038	-0.003	-0.651	-0.362	-0.012
time	-0.024	-0.003	0.003	0.000	-0.056	0.000
time × treatment	-0.006	0.000	-0.001	-0.184	-0.119	0.007
intercept	-0.156	0.557	0.034	0.104	1.692	0.053
treatment	-0.067	0.068	-0.006	-0.661	-0.346	-0.011
time	-0.519	-0.075	0.070	0.001	-0.424	0.000
time × treatment	-0.084	-0.004	-0.018	-0.990	-0.593	0.037

We see especially from the lower part of Table 1 that INLA shows large relative differences compared to MCMC but also to ML. While the relative differences in the RI model are not larger than 0.519, the differences are substantially larger in the RI+RS model with values up to 0.99. Relative differences also increase if we compare INLA with ML, to a maximum of 0.557 for the RI model and 1.692 for the RI+RS model. In contrast, the estimates based on MCMC are much closer to the ML estimates, with a maximum relative difference of 0.07.

The differences shown in the upper half of Table 1 for the RI-model may be considered as acceptable, with a maximal difference of 0.073 on the log-odds ratio scale. However, more substantial discrepancies can be seen for the RI+RS model, in particular for comparisons involving INLA estimates. See Table 1 in Appendix B for the fixed effect estimates of the models presented in Figure 1, 2 and Table 1.

The argument `inla.control` includes several settings which can be modified and which affect the accuracy of the numerical integration of the hyperparameters in `r-inla`. We increased the numerical accuracy and set the step length for the integration to `dz = 0.2` from the default value `dz = 1`, the step length for the gradient calculations to `h = 1e-5` from default `h = 0.01`, the tolerance criteria for the change in the posterior to `tolerance = 1e-6` from default `tolerance = 0.005` and we changed the integration strategy to `int.strategy = "grid"` which uses as default the less accurate central composite design (`int.strategy = "ccd"`). The differences between the posterior distributions shown in Figure 1 and 2 only improved slightly by using these settings, compared to the default ones.

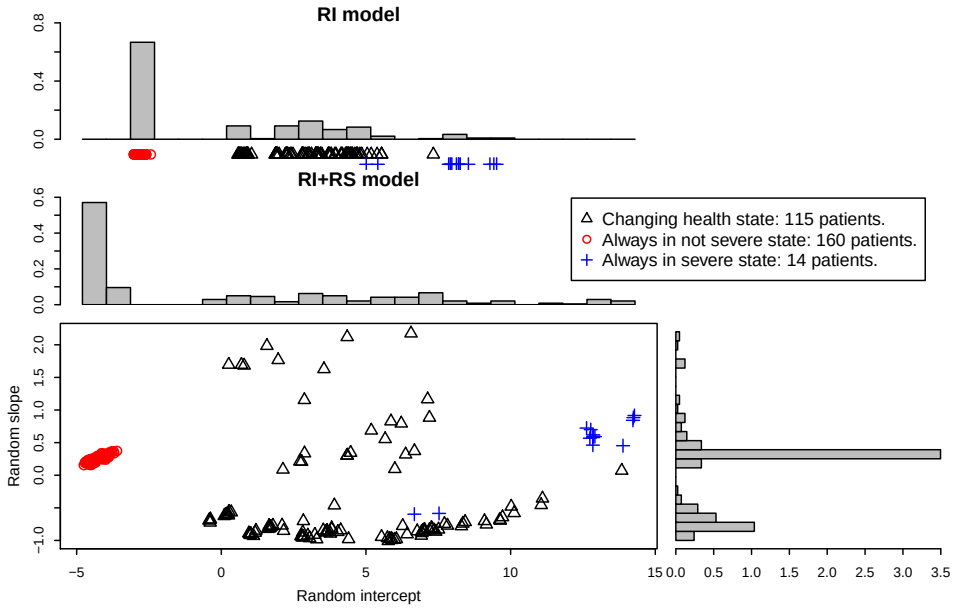

Figure 3. Random effect estimates $\mathbf{b}_i$ for the RI and the RI+RS model. Estimates are marginal posterior means obtained by MCMC.

Throughout the paper we used the default `simplified.laplace` approximation strategy in `r-inla`. Changing the approximation strategy to `laplace` did not reduce the differences for the posterior distributions as illustrated in the Supplementary Material. Additionally, Figure 2 in the Supplementary Material shows the marginal posterior distributions for the hyperparameters, which are substantially different for INLA compared to ones based on MCMC.

3.2. *Cluster-specific quasi-complete separation*

Table 1 in Section 3.1 clearly indicates that differences between MCMC and INLA, relative to the MCMC standard deviation, exceed the previously reported 30% for binary response GLMMs [2]. A correction in the location of the posterior distribution has been recommended as a possible error-correction [1]. But none of the different approximation strategies did improve the location shift of the marginal posteriors obtained by INLA. The differences between INLA and MCMC got even more pronounced if the time variable was not centred.

A closer look at the random effect estimates, obtained by MCMC, gives some interesting details. Figure 3 gives histograms of the means of the marginal posterior distribution for the random effects. The upper part shows the random effect estimates for the RI and the lower part for the RI+RS model. An additional scatter plot gives the joint distribution of estimated random intercepts and slopes in the RI+RS model. Three clusters can be distinguished: there are 160 patients, who always stay in the non severe state during the observation period (marked with a red circle), 14 patients who stay always in the severe state (marked with a blue cross), while the remaining 115 patients (marked with black triangles) switch their health state at least once. Figure 3 indicates that patients without any variation in the response build clusters and take extreme values for the random effect estimates. As a result, the empirical distribution of the random effect

estimates does not resemble a normal distribution. This hints to a substantial cluster-specific quasi-complete separation problem for the toenail data, as discussed in Section 2.3.

However, there are two patients who are always in the severe response category but their random intercept does not cluster with random effects from the other patients who always stay in the severe state. The reason for the comparably low random intercept is that these two patients are only observed at two, respective three follow-up visits. Thus they are not close to the patients who were observed seven times in the severe response state. Also they were only observed at centered times below zero such that their random slope estimate is negative. On the other hand there is one random effect which is close to the cluster of random effects for patients always being in the same state, although this patient switches the response. This patient was observed at seven occasions but only at the very last observation a moderate infection was declared, such that this profile is very similar to having always a response equal to one.

4. Simulation with varying cluster-specific quasi-complete separation

To assess a possible problem of INLA with cluster-specific quasi-complete separation in more detail, we undertook two simulation studies, with a varying proportion of cluster-specific quasi-complete separation. The first study is based on one simulated dataset only, for which we manipulate the response such that the proportion of patients always having response equal to zero changes. The results are used to examine if the differences between INLA and MCMC respectively ML, discussed in Section 3.1, are persistent or just a random artefact of the toenail data. In the second simulation study we investigate the accuracy of the parameter estimates by INLA by randomly generating replicates of the dataset and assessing the root mean squared error and bias.

In Section 4.1 and 4.2 we simulate balanced datasets with n observations per patient otherwise similar to the toenail infection data. The observed time period ranges from -4.5 to 4.5 and the time differences between follow-up visits are rescaled according to the choice of n. We set the fixed effect for time to 0.8 and for the time treatment interaction to -0.8 while the main effect for treatment is assumed to be zero. The random intercept standard deviation σ_b is set to two.

4.1. *Comparison of estimation methods by changing the response*

For the comparison of the three estimation methods we generated one initial dataset, with $I = 300$ patients observed at baseline and six follow-up visits ($n = 7$). The simulated dataset was guaranteed to have an initial proportion of patients with constant response profile fixed at 5% of all patients. We then successively started to manipulate the response of this dataset and increased the proportion of patients who always remain in state zero by another 5% or 15 patients, until we reached a proportion of 75%. In this way we continually increased the degree of cluster-specific quasi-complete separation in the data.

The plots in Figure 4 show the absolute differences between the parameter estimates by INLA, MCMC and ML. As before, the comparison with ML is based on fixed hyper-parameters. The upper half in Figure 4 shows these differences for a RI and the lower part for a RI+RS model. We see that INLA and MCMC agree well if there is only a low proportion of patients who always remain in the same health state. But with increasing proportion of patients with a constant response profile, the differences between MCMC and INLA increase. If the proportion is 40% or larger, we see substantial discrepancies. The same pattern is visible in the middle panel of Figure 4, where INLA is compared

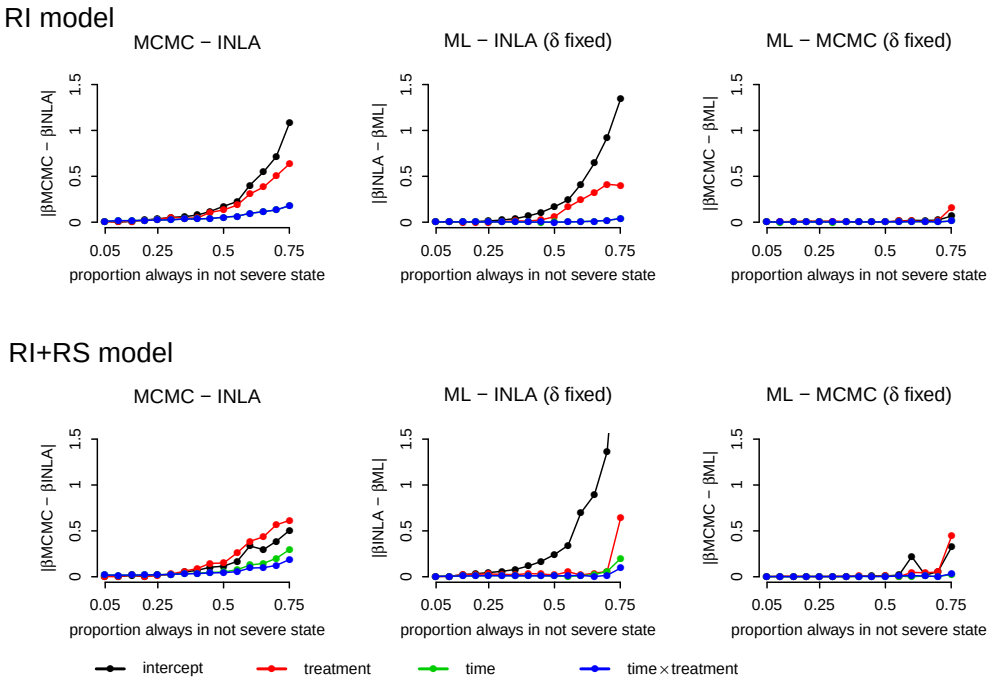

Figure 4. Absolute differences between the marginal posterior means obtained with INLA and MCMC and with ML estimates for simulated data with increasing degree of cluster-specific quasi-complete separation for a RI (upper row) and RI+RS (bottom row) model. Hyperparameters were fixed at the ML estimates for comparisons of INLA and MCMC with ML. The value for the absolute difference of the intercept between ML and INLA (δ fixed) with a proportion of 75% of patients always in the not severe state is 3.1 and not shown in the corresponding plot.

with ML. Here the differences for the fixed intercept seem to increase even more quickly. The last column in Figure 4 compares MCMC with ML which does not show any large differences, even for a large proportion of patients with a constant response profile. The few occasional differences between MCMC and ML for the RI+RS model, shown in the bottom right plot of Figure 4, may be explained by the unstable fixed effect parameter estimates based on `lme4` (see Appendix A). In this simulation we always used 40 quadrature points in the GHQ-approximation.

4.2. *Assessment of root mean squared error and bias*

The comparison of methods in Section 4.1 shows that discrepancies between INLA and the other two methods increase along with increasing degree of cluster-specific quasi-complete separation. Therefore we assess the accuracy of INLA estimates in the following simulation study. In order to keep the scope limited we report results for the random intercept model only. There are three parameters which we allow to vary, the number of patients I, the number of observations per patient n and the fixed intercept. We used four different settings with I equal to 50 or 125 and with n equal to 10 or 25. The fixed intercept is varying from -8.5 to -2 by 0.5 steps such that the proportion of patients always observed in the same state is varying. The fixed intercepts were chosen such that a large range of cluster-specific quasi-complete separation results in the simulated datasets. For a given fixed intercept the proportion of patients always observed in the same state is not necessarily the same across the four different scenarios. Still the range of cluster-specific

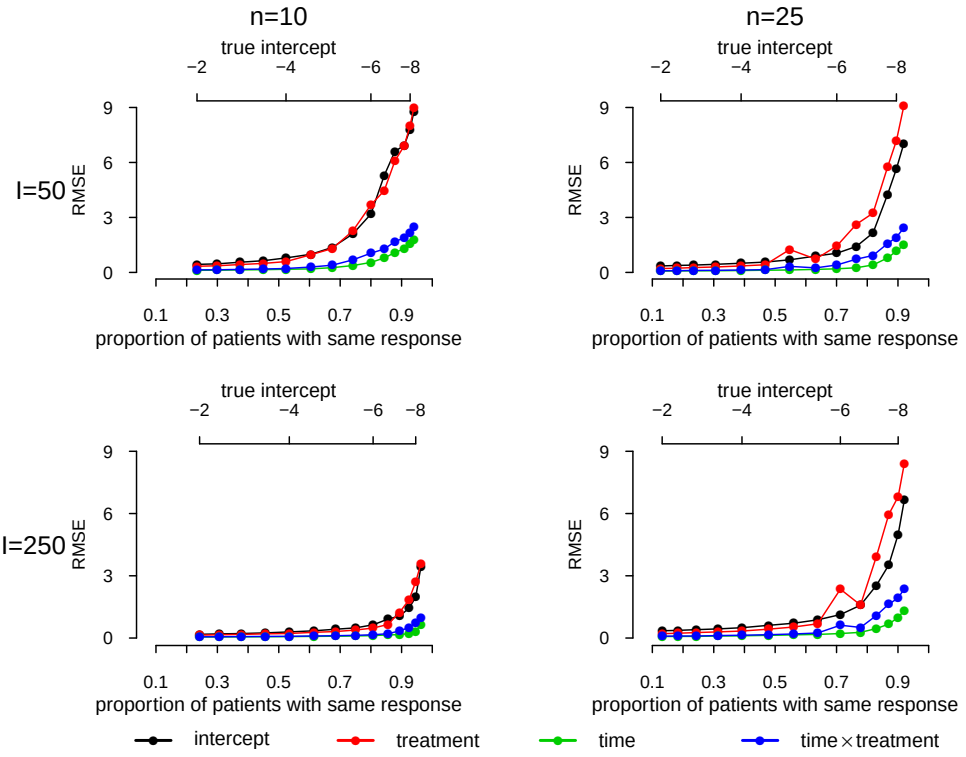

Figure 5. RMSE for marginal posterior means of fixed effects $\boldsymbol{\beta}$ for a RI model based on 1'000 iterations with different number of patients I and number of observations per patient n.

quasi-complete separation reaches from below 20% to above 80% in all four simulations.

For each of the 14 different intercepts and for each of the four scenarios we iteratively simulate 1'000 datasets, resulting in 56'000 `r-inla` calls. We rule out any datasets which include a complete separation of the response given the treatment, which would result in diverging fixed effect estimates and repeat the iteration if this occurs. For the four parameter combinations we report the root mean squared error RMSE= $\sqrt{1/N \sum_{i=1}^{N} (\hat{\beta}_i - \beta)^2}$ in Figure 5 and the bias $1/N \sum_{i=1}^{N} \hat{\beta}_i - \beta$ in Figure 6 based on the marginal posterior means.

Figure 5 shows for all four combinations of I and n an increasing RMSE with increasing proportion of patients always having the same response for all fixed effect estimates. Although the simulation with $n = 10$ and $I = 125$, in the bottom left plot, has a lower RMSE compared to the other three scenarios. The RMSE in all plots of Figure 5 is increasing with increasing quasi-complete separation Figure 6 illustrates that there is also an increasing bias with increasing proportion of quasi-complete separation. Compared to the other three scenarios Figure 6 shows that for $I = 125$ and $n = 10$ the assessed bias is relatively small.

Although the intercept has the largest bias also the other fixed effects are affected increasingly by increasing cluster-specific quasi-complete separation. This is in line with the results for the toenail dataset in Section 3.1 and also with the simulation in Section 4.1, where the estimates for the intercept based on INLA had the largest difference compared to the other two methods.

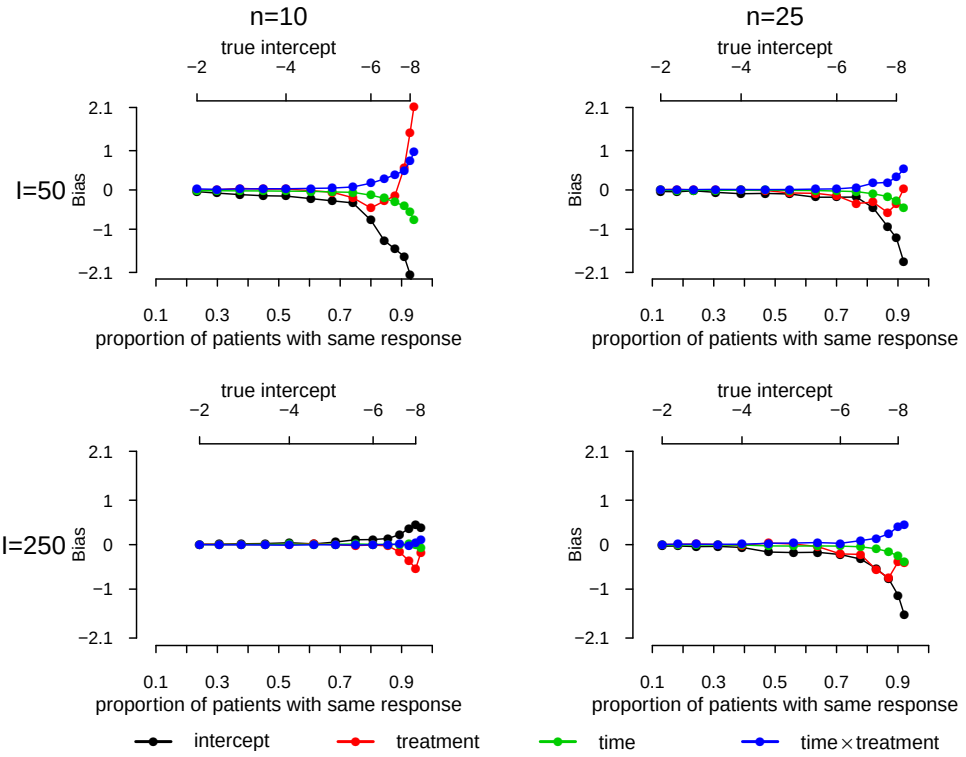

Figure 6. Bias for marginal posterior means of fixed effects $\boldsymbol{\beta}$ for a RI model based on 1'000 iterations with different number of patients I and number of observations per patient n.

5. Discussion

We showed that the approximation error by INLA increases for binary response GLMMs if the data shows a substantial and increasing degree of cluster-specific quasi-complete separation. INLA estimates agree rather well with MCMC and ML, unless the degree of cluster-specific quasi-complete separation is high. The simulation in Section 4.1 disclosed already large discrepancies if the proportion of patients with a constant response is 40%. Differences shown in Figure 4 are in the same range as the ones found for the toenail infection trial in Section 3.1, where 55.4% of the patients always stayed in the not severe state. This large degree of cluster-specific quasi-complete separation causes INLA to fail to produce reliable parameter estimates.

This was confirmed by the simulation study in Section 4.2, which illustrated that the RMSE as well as the bias increases with increasing proportion of patients always being in the same state. Although MCMC sampling is known to converge to the true posterior distribution if the number of samples is large enough, it would require much more computing time to analyse the same number of replicated datasets. However, INLA is much faster than MCMC, required less than 10 seconds per call and thus it was possible to assess the RMSE and bias with `r-inla` based on this rather large number of replicates with modest computational effort.

As illustrated in Appendix A also ML estimation may result in numerical instabilities in such situations and MCMC may request a large number of iterations. However, only INLA shows already at a comparably low degree of cluster-specific quasi-complete sep-

aration a systematic bias. This finding contrasts the results by [2] and [3], who do not investigate this scenario and thus are too optimistic regarding INLA's accuracy.

In the context of Bayesian inference most often critique is directed to the choice of the prior distributions. Usually one would assume that there must be a possibly very informative prior, which helps to stabilize the deteriorating INLA estimates if cluster-specific quasi-complete separation is present. We thus looked at different prior specifications for the hyperparameters. It has been argued [23] that an inverse gamma prior on the random effects variances may result in large sensitivity of parameter estimates. Indeed, the alternative half-normal prior distribution on σ_b [23] shows less prior sensitivity [32]. We therefore investigated if part of the discrepancies between INLA and MCMC are due to the inverse gamma prior in the RI model. Naturally, as consequence of adapting the prior, the parameter estimates changed for the toenail data. However, the differences between INLA and MCMC did not decrease, such that our main findings persisted under the alternative half-normal, and also under more informative prior specifications. Another model modification is to relax the normality assumption for the random effects in the RI model and to use a t-distribution. This model can be considered in `r-inla` [33], but differences still did not decrease substantially.

Alternatively, the non-normal distribution of the random effects, shown in Figure 3, suggests to use a mixture of normal distributions [34, 35]. This formulation has been shown to provide a better fit to the data [36]. However, implementation of such a mixture model in INLA is not straightforward and a combination with an expectation-maximization (EM) type algorithm might be required [37].

Nevertheless there are possibly ways in how this specific problem could be addressed in INLA to improve its performance, *e. g.* in [1] section 6.1 a possible alternative way to approximate the posterior marginals for the hyperparameters based on a Gaussian copula is mentioned. Finally it is important to highlight that (quasi) complete separation in mixed models is not INLA related, but a general problem, for which awareness should be high, indifferently what kind of inference is applied. If encountering cluster-specific quasi-complete separation for a binary response GLMM based on longitudinal data, one could perhaps avoid this by Markov models based on time-dependent transition probabilities $\Pr(y_{ij} = 1 \,|\, y_{i(j-1)}, \boldsymbol{\beta}, \mathbf{b}_i, \mathbf{D})$ instead of $\Pr(y_{ij} = 1 \,|\, \boldsymbol{\beta}, \mathbf{b}_i, \mathbf{D})$ as discussed in [27, 38].

If using INLA for a binary, or even binomial GLMMs, one should always pay some effort in investigating if there is cluster-specific quasi-complete separation present in the data. In practice one should check if the variance for the hyperparameters is large and if there are clusters with very high and very low random effect estimates. These may be valuable hints towards a possible large cluster-specific quasi-complete separation, which requires further investigation, as INLA may under these circumstances provide biased parameter estimates.

Acknowledgements

A lot of support and insights resulted from discussions with Andrea Riebler and Håvard Rue. Comments and suggestions of two anonymous reviewers helped to substantially improve the paper.

References

[1] Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). Journal of the Royal Statistical Society - Series B. 2009 2;71:319–392.

[2] Fong Y, Rue H, Wakefield J. Bayesian inference for generalized linear mixed models. Biostatistics. 2010;11(3):397–412.
[3] Grilli L, Metelli S, Rampichini C. Bayesian estimation with integrated nested Laplace approximation for binary logit mixed models. Journal of Statistical Computation and Simulation. 2014 July;0(0):1–9.
[4] Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White JSS. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology & Evolution. 2009;24(3):127 – 135.
[5] Zhang H, Lu N, Feng C, Thurston SW, Xia Y, Zhu L, Tu XM. On fitting generalized linear mixed-effects models for binary responses using different statistical packages. Statistics in Medicine. 2011;30(20):2562–2572.
[6] Capanu M, Gönen M, Begg CB. An assessment of estimation methods for generalized linear mixed models with binary outcomes. Statistics in Medicine. 2013;32(26):4550–4566.
[7] Minka T. A family of algorithms for approximate Bayesian inference [dissertation]. MIT; 2001.
[8] Kuss M, Rasmussen CE, Herbrich R. Assessing approximate inference for binary Gaussian process classification. Journal of Machine Learning Research. 2005;6:1679–1704.
[9] Albert A, Anderson JA. On the existence of maximum likelihood estimates in logistic regression models. Biometrika. 1984;71(1):1–10.
[10] Tierney L, Kadane JB. Accurate approximations for posterior moments and marginal densities. Journal of the American Statistical Association. 1986;81(393):82–86.
[11] Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association. 1993;88(421):9–25.
[12] Breslow NE, Lin X. Bias correction in generalised linear mixed models with a single component of dispersion. Biometrika. 1995;82(1):81–91.
[13] Lin X, Breslow NE. Bias correction in generalized linear mixed models with multiple components of dispersion. Journal of the American Statistical Association. 1996;91(435):1007–1016.
[14] Pinheiro JC, Bates DM. Approximations to the log-likelihood function in the nonlinear mixed-effects model. Journal of Computational and Graphical Statistics. 1995;4(1):12–35.
[15] Rue H, Held L. Gaussian Markov Random Fields: Theory and Applications. London: Chapman & Hall/CRC Press; 2005.
[16] Gamerman D. Sampling from the posterior distribution in generalized linear mixed models. Statistics and Computing. 1997 MAR;7(1):57–68.
[17] Holmes CC, Held L. Bayesian auxiliary variable models for binary and multinomial regression. Bayesian Analysis. 2006;1(1):145–168.
[18] Held L, Sabanés Bové D. Applied Statistical Inference; Likelihood and Bayes. Springer; 2014.
[19] Firth D. Bias reduction of maximum likelihood estimates. Biometrika. 1993;80(1):27–38.
[20] Heinze G, Schemper M. A solution to the problem of separation in logistic regression. Statistics in Medicine. 2002;21(16):2409–2419.
[21] Heinze G. A comparative investigation of methods for logistic regression with separated or nearly separated data. Statistics in Medicine. 2006;25(24):4216–4226.
[22] Abrahantes JC, Aerts M. A solution to separation for clustered binary data. Statistical Modelling. 2012;12(1):3–27.
[23] Gelman A, Jakulin A, Pittau MG, Su YS. A weakly informative default prior distribution for logistic and other regression models. Annals of Applied Statistics. 2008;2(4):1360–1383.
[24] Gelman A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian Analysis. 2006 09;1(3):515–534.
[25] De Backer M, De Vroey C, Lesaffre E, Scheys I, De Keyser P. Twelve weeks of continuous oral therapy for toenail onychomycosis caused by dermatophytes: A double-blind comparative trial of terbinafine 250 mg/day versus itraconazole 200 mg/day. Journal of the American Academy of Dermatology. 1998;38(5, Supplement 2):S57 – S63.
[26] Lesaffre E, Spiessens B. On the effect of the number of quadrature points in a logistic random effects model: an example. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2001;50(3):325–335.
[27] Molenberghs G, Verbeke G. Models for Discrete Longitudinal Data. Springer; 2005.
[28] Plummer M. JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs

Sampling. In: Proceedings of the 3rd International Workshop on Distributed Statistical Computing; 2003.
[29] Plummer M, Best N, Cowles K, Vines K. CODA: convergence diagnosis and output analysis for MCMC. R News. 2006;6(1):7–11.
[30] Frühwirth-Schnatter S, Frühwirth R, Held L, Rue H. Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data. Statistics and Computing. 2009;19(4):479–492.
[31] Bates D, Maechler M, Bolker B. lme4: Linear mixed-effects models using S4 classes.. 2011.
[32] Roos M, Held L. Sensitivity analysis in Bayesian generalized linear mixed models for binary data. Bayesian Analysis. 2011;6(2):259–278.
[33] Martins TG, Rue H. Extending integrated nested Laplace approximation to a class of near-Gaussian latent models. Scandinavian Journal of Statistics. 2014;:online.
[34] Verbeke G, Lesaffre E. A linear mixed-effects model with heterogeneity in the random-effects population. Journal of the American Statistical Association. 1996;91(433):217–221.
[35] Komárek A, Lesaffre E. Generalized linear mixed model with a penalized Gaussian mixture as a random effects distribution. Computational Statistics & Data Analysis. 2008;52(7):3441 – 3458.
[36] Verbeke G, Molenberghs G. The gradient function as an exploratory goodness-of-fit assessment of the random-effects distribution in mixed models. Biostatistics. 2013;14(3):477–490.
[37] Van De Wiel MA, Leday GG, Pardo L, Rue H, Van Der Vaart AW, Van Wieringen WN. Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. Biostatistics. 2013;14(1):113–128.
[38] Diggle PJ, Heagerty PJ, Liang KY, Zeger SL. Analysis of Longitudinal Data. 2nd ed. Oxford Statistical Science Series; Oxford: Oxford University Press; 2003.

Appendix A. ML estimation for toenail data and varying quadrature points

The choice of the number of quadrature points may have an influence on the fixed effect estimates in RI and RI+RS models [26]. Specific implementations may differ in different software packages [5]. We illustrate in Figure A1 that the fixed effect estimates for the toenail data are varying with the number of quadrature points used in the adaptive GHQ-approximation. This confirms the findings by [26] who compared adaptive and non-adaptive GHQ-approximation. Figure A1 shows differences for the fixed effect estimates obtained by `PROC NLMIXED` in `SAS` and `lme4` in `R` confirming the findings by [5] who also state a large difference between the two software implementations. Figure A1 suggests that, to obtain accurate estimates, the RI+RS model needs more quadrature points than the simpler RI model. Strikingly, the fixed effects obtained by `lme4` start to vary again for more than 81 quadrature points. Additionally `lme4` repeatedly produced a warning message resulting from the optimization algorithm `nlminb` which is indicated by small bars at the bottom of the two `glmer` plots. For 82 and 83 quadrature points, indicated with crosses, `glmer` aborted with an error message.

Again the `lme4` version 0.999999-2 was used, as the number of quadrature points is hard coded to a maximum of 25 in later versions. The `R` version 2.15.3 (2013-03-01) was used. If a newer `R` version together with `lme4` version 0.999999-2 is used, convergence criteria and related error and warning messages may be different. For `SAS` we used version 9.3. In contrast to the text above, models shown in Figure A1 are based on the data with uncentred timescale. Due to randomization of the trial and the uncentred timescale the treatment effect was omitted for the models here.

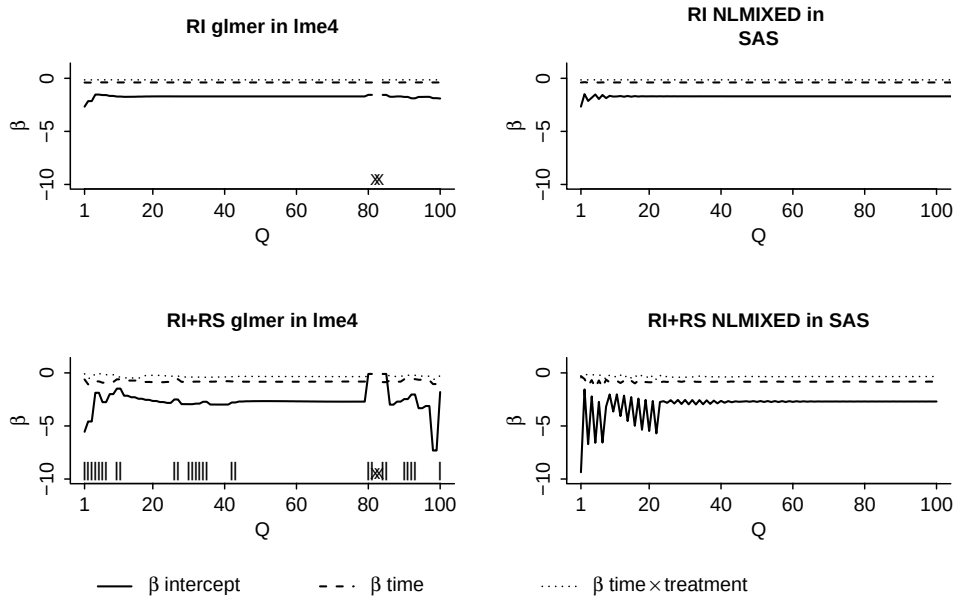

Figure A1. Fixed effects estimates for toenail data with varying number of quadrature points. A x indicates that `glmer` finished with "Error message: Downdated X'X is not positive definite, 1." and | that `glmer` finished with "Warning message:In mer_finalize(ans) : false convergence (8)".

Appendix B. Parameter estimates for RI and RI+RS model for toenail data

Table B1. Fixed effect estimates and hyperparameters by INLA, MCMC and ML estimation for the RI (upper part) and RI+RS model (lower part). For the fixed effects the standard errors are shown in parentheses. For INLA and MCMC the means of the marginal posterior distribution are shown. The last two columns show the results if the hyperparameter values are fixed at the estimates obtained by ML.

	INLA	MCMC	ML	INLA-fix	MCMC-fix
intercept	-3.441	-3.515	-3.482	-3.707	-3.495
	(0.406)	(0.469)	(0.396)	(0.386)	(0.404)
treatment	-0.737	-0.778	-0.753	-0.791	-0.749
	(0.507)	(0.605)	(0.571)	(0.552)	(0.566)
time	-0.372	-0.396	-0.390	-0.387	-0.393
	(0.0404)	(0.0460)	(0.0434)	(0.0402)	(0.0435)
time × treatment	-0.133	-0.139	-0.139	-0.139	-0.138
	(0.0618)	(0.0705)	(0.0709)	(0.0640)	(0.0696)
σ_b^2	12.848	16.858	16.036		
intercept	-6.280	-6.180	-6.588	-7.816	-6.627
	(0.797)	(0.974)	(0.766)	(0.703)	(0.726)
treatment	-0.848	-1.499	-1.593	-1.231	-1.581
	(0.818)	(0.985)	(1.144)	(1.039)	(1.047)
time	-0.761	-0.761	-0.824	-0.768	-0.824
	(0.147)	(0.175)	(0.151)	(0.136)	(0.132)
time × treatment	-0.144	-0.328	-0.344	-0.224	-0.351
	(0.184)	(0.186)	(0.239)	(0.211)	(0.201)
$\sigma_{b_1}^2$	25.7083	42.7380	47.7495		
$\sigma_{b_2}^2$	0.7441	0.9055	1.0356		
ρ	0.0249	-0.0742	-0.0531		

Supplementary material to "Quasi-complete Separation in Random Effects of Binary Response Mixed Models"

Rafael Sauter and Leonhard Held
Department of Biostatistics
Epidemiology, Biostatistics and Prevention Institute
University of Zurich
Hirschengraben 84, 8001 Zurich, Switzerland
Email: rafael.sauter@uzh.ch, leonhard.held@uzh.ch

24th February 2016

This supplementary material provides additional information to the main text of the paper "Quasi-complete Separation in Random Effects of Binary Response Mixed Models". The supplementary material contains additional information to the random intercept (RI) and the random intercept plus random slope (RI+RS) models, presented in the main text in Section 3. In Section 1 the marginal posterior distributions for the fixed effects and the hyperparameters obtained with integrated nested Laplace approximations (INLA) with the simplified Laplace approximation approach and with the full Laplace approximation are compared. Section 2 provides convergence diagnostics for the models parameters obtained with MCMC.

1 Simplified and full Laplace approximations in INLA

The following plots compare the marginal posterior distributions of the fixed effect parameters and hyperparameters obtained with INLA and MCMC. For the simplified Laplace approximation the `inla.control` option is set to

```
control.inla = list(
  strategy = "simplified.laplace",
  int.strategy = "grid",
  h = 1e-5,
  tolerance = 1e-6)}
```

which are the settings used also for the results reported in the main text. For the full Laplace approximation the `inla.control` option is set to

```
control.inla = list(
                  strategy = "laplace",
                  fast = FALSE,
                  int.strategy="grid",
                  h = 1e-5,
                  tolerance = 1e-6)
```

for which the results are reported in the following Figure 1 to 3. In general the full Laplace approximation in INLA is more accurate compared to the simplified Laplace approximation but in most cases nearly coincide according to **?**. However, for the RI and RI+RS models of the toenail infection data we found substantial differences between the two approximation strategies and the simplified Laplace approximation was closer to the results obtained by MCMC, if the hyperparameters are not fixed. This differences may well be related to the problem of substantial cluster-specific quasi-complete separation. In the main text the reported results are always based on the simplified Laplace approximation.

1.1 Estimation with prior distribution

In Figure 1 the marginal posterior distributions for the fixed effects β based on INLA with simplified and with full Laplace approximation and with MCMC are shown with a prior distribution on the hyperparameters. The same prior distributions for the hyperparameters as described in the main text are used. Figure 2 shows the corresponding marginal posterior distributions for the hyperparameters.

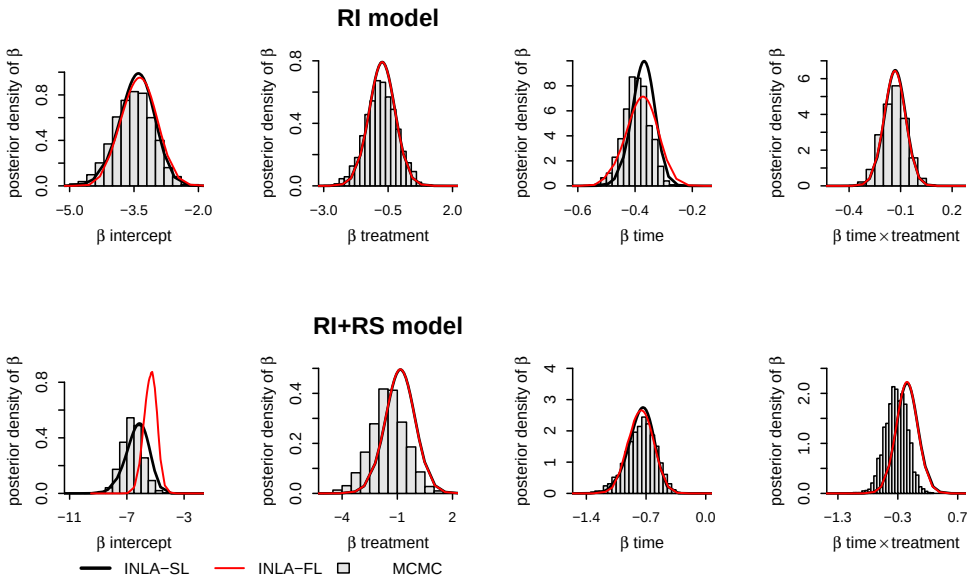

Figure 1: Marginal posterior distributions of β by MCMC (histogram), INLA with simplified Laplace (INLA-SL, black line) and based on INLA with full Laplace (INLA-FL, red line).

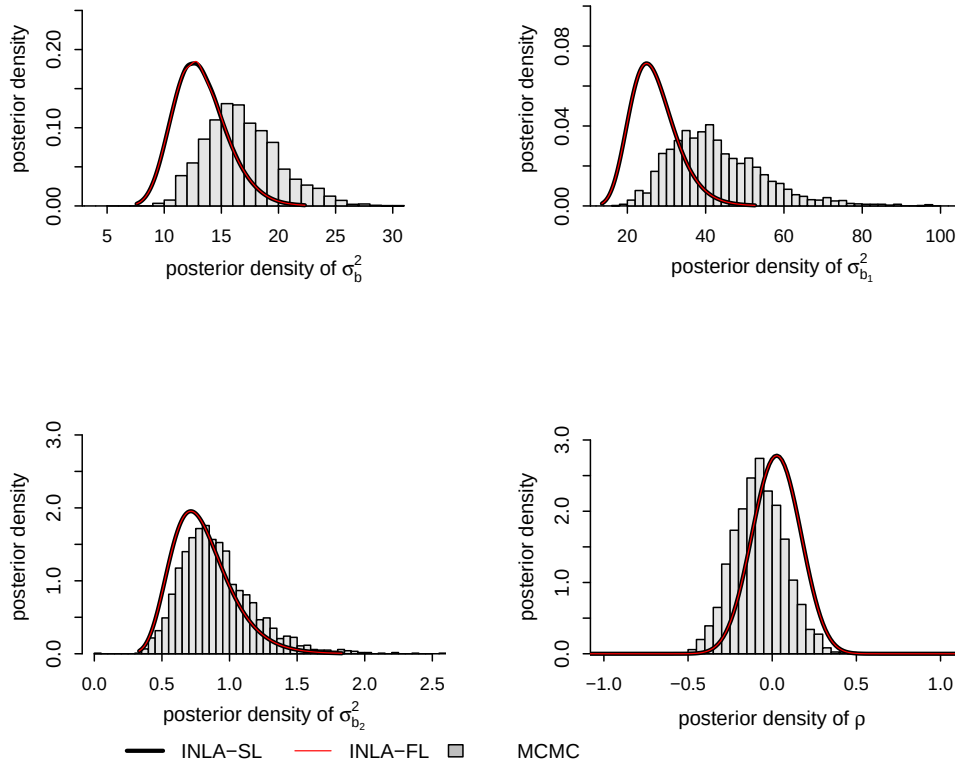

Figure 2: Marginal posterior distributions of hyperparameters for the RI and RI+RS models by MCMC (histogram), INLA with simplified Laplace (INLA-SL, black line) and based on INLA with full Laplace (INLA-FL, red line).

1.2 Estimation with fixed hyperparameters

In Figure 3 the marginal posterior distributions for the fixed effects $\boldsymbol{\beta}$ based on INLA with simplified and with full Laplace approximation and with MCMC are shown with fixed hyperparameters. The hyperparameters are fixed at the same values as obtained by ML estimation and as indicated in the main text.

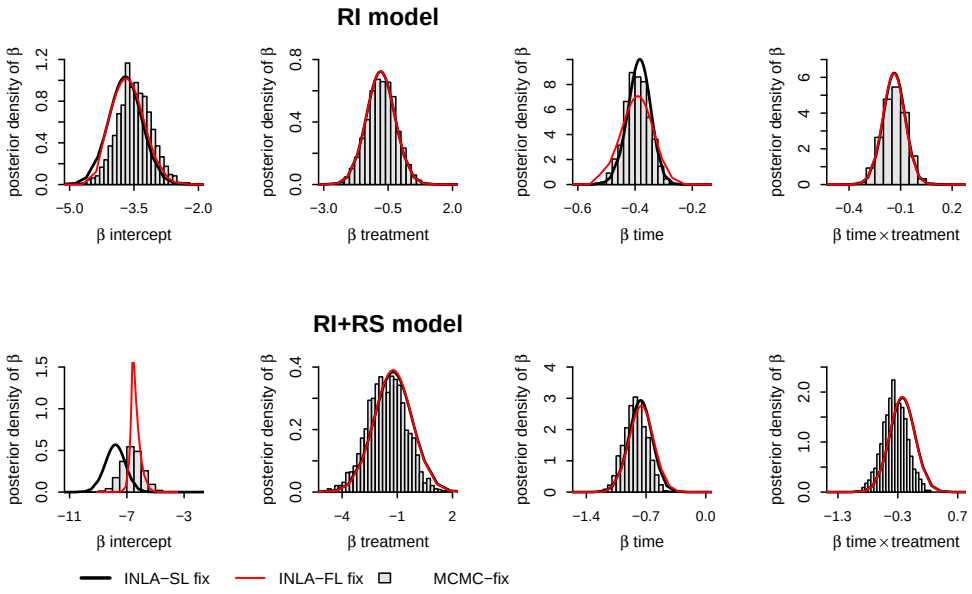

Figure 3: Marginal posterior distributions of $\boldsymbol{\beta}$ by MCMC (histogram), INLA with simplified Laplace (INLA-SL fix, black line) and based on INLA with full Laplace (INLA-FL fix, red line). Hyperparameters values are fixed at corresponding ML estimates.

2 Convergence diagnostics

In this section we provide convergence diagnostics of the MCMC run for the RI and RI+RS model based on the toenail dataset for which the results were presented in Section 3 in the main text. Convergence diagnostics are reported for all four models, the RI and the RI+RS model with a prior on the hyperparameters and with fixed hyperparameters. We show the convergence diagnostics for all fixed effects ($\boldsymbol{\beta}$) of each model, of the hyperparameters if not fixed and for the random effect estimates of patient 233, who is one of the patients which always had a response value equal to one, plus two additional random effects of two patients in each model with the two highest autocorrelations at lag 1 in the MCMC run. For each reported parameter we show a excerpt of the traceplot, the autocorrelation as well as Geweke convergence diagnostics.

The excerpt for the traceplot covers the last 500 iterations of the MCMC run from iteration 1900 to 2400. We only report part of the MCMC run in the traceplot as the plot based on all 2400 kept iterations does not show any details about structures in the traceplot. Before reporting this excerpt, each traceplot based on all iterations was inspected to have possible jumps in the Markov chain. This was not the case for none of the reported parameters of the four models such that the shown excerpt is representative for the complete traceplot and thus the mixing of the Markov chain was found to be sufficient.

The autocorrelation was determined by the function `autcorr.diag` in the `coda` package and the maximum lag length was set to 10. The index in the autcorrelation plots and the Geweke diagnostic are values without thinning *e. g.* a autocorrelation for lag 1 has an index equal to iteration 200 in the plot. Except for the intercept, the time, the time treatment interaction and the hyperparameters in the RI+RS, there is no substantial autocorrelation in the MCMC run of the fixed effect parameters. Even for these coefficients in the RI+RS model the autocorrelation is modest and drops quickly to zero after six lags and if hyperparameters are fixed in the RI+RS model there is no autocorrelation present any more.

The Geweke diagnostics are computed by the function `geweke.plot` in the `coda` package in `R`. The Geweke statistic compares the posterior means computed based on the second half of the MCMC iterations with the means computed based on a decreas-

ing fraction of the first half of the MCMC iterations. The Geweke test statistic is a standardized z-score based on the difference between the two sample means divided by its estimated standard error. The standard error is estimated from the spectral density at zero and so takes into account any autocorrelation. The plot with the Geweke diagnostics shows what happens if successively larger numbers of iterations are discarded from the beginning of the chain. The first half of the Markov chain is divided into 7 segments, for which the Geweke's Z-score is repeatedly calculated. The first Z-score is calculated with all iterations in the chain, the second after discarding the first segment, the third after discarding the first two segments, and so on. The last Z-score is calculated using only the samples in the second half of the chain. The two horizontal dashed lines indicated the 2.5% and 97.5% quantile of the standard normal distribution. A Geweke Z-score which is far away from the dashed lines implies that the two means, of the first and the second part of the Markov chain, are probably not equal which implies that the stationary distribution was not reached. The presented plots with the Geweke statistics show that some of the Geweke Z-scores are outside the 2.5% or the 97.5% interval but all of them are still close to these quantiles and all of them have absolute values which are smaller than three.

From this analysis we would finally not reject the hypothesis of convergence to a stationary distribution of the MCMC run for the presented models. This was clearly not the case for MCMC runs with a smaller number of iterations, less thinning or based on different MCMC samplers.

2.1 Random intercept model (RI) with prior

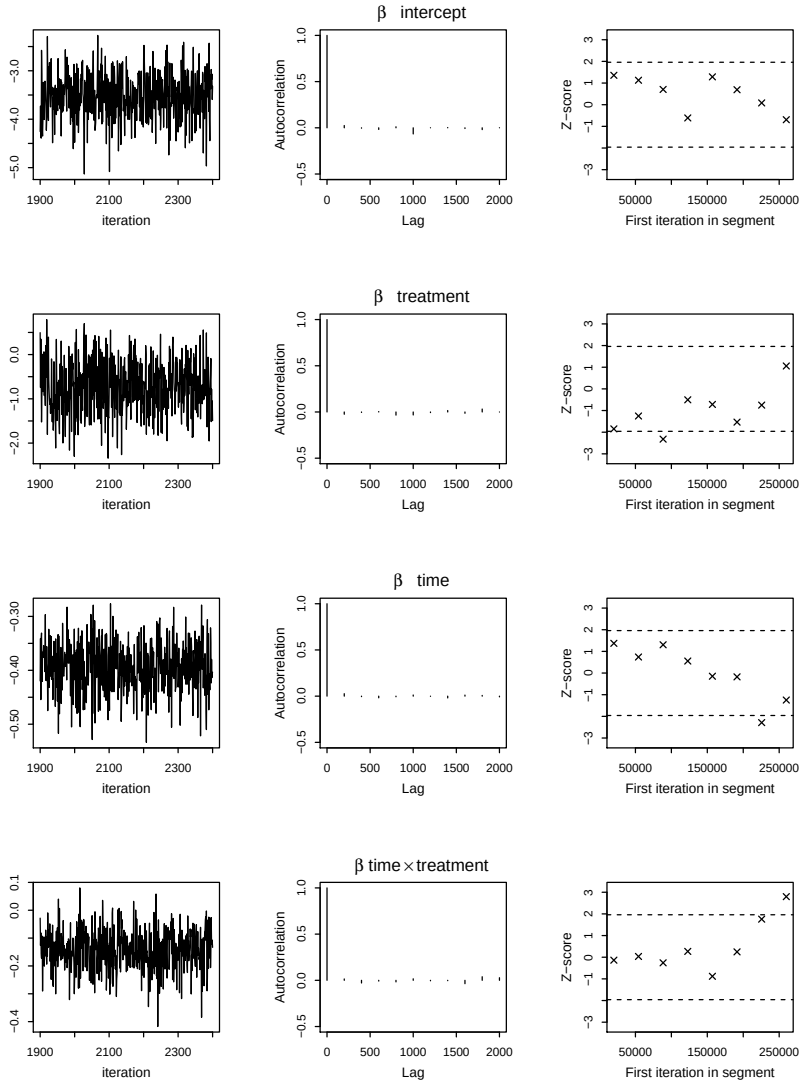

Figure 4: Convergence diagnostics for fixed effects of RI model: traceplot, autocorrelation and Geweke diagnostics.

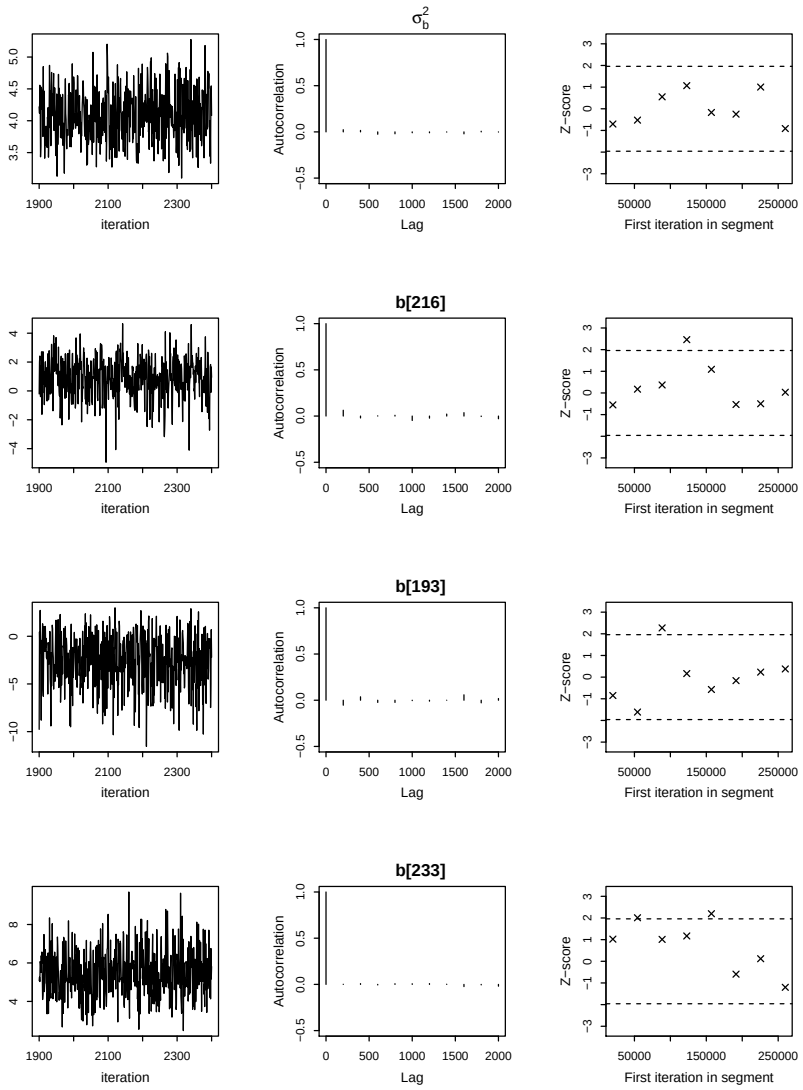

Figure 5: Convergence diagnostics for hyperparameters and selected random effects of RI model: traceplot, autocorrelation and Geweke diagnostics.

2.2 Random intercept model (RI) with fixed hyperparameters

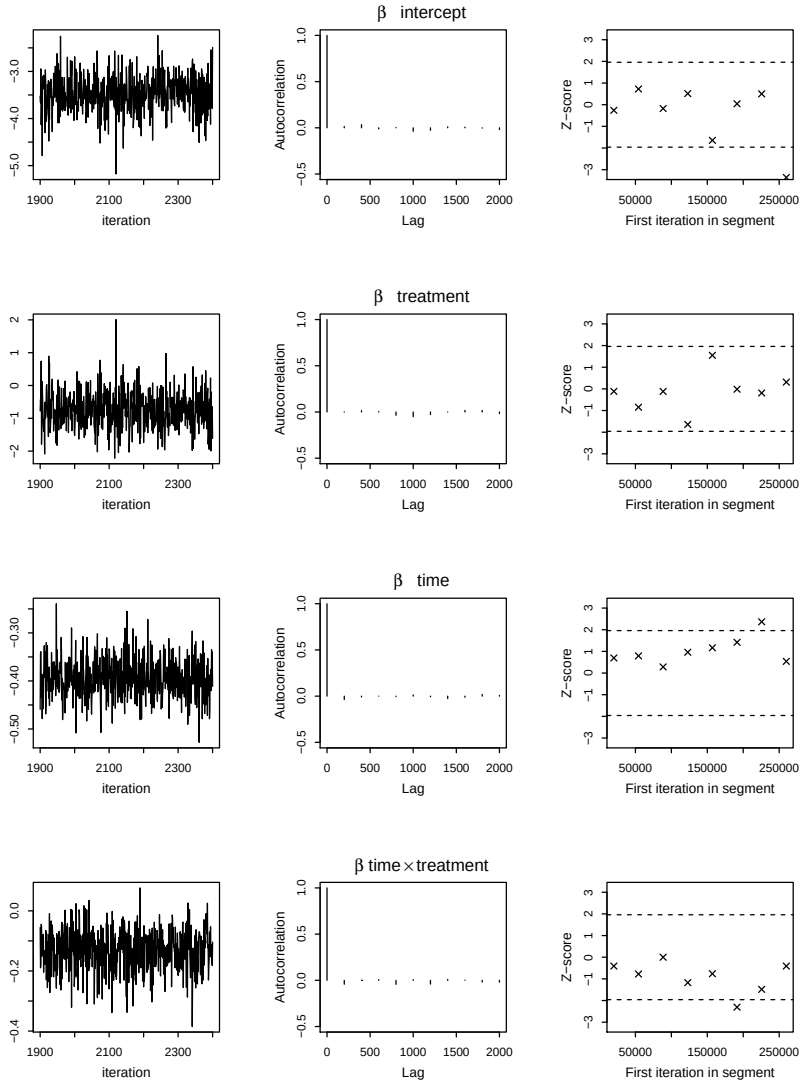

Figure 6: Convergence diagnostics for fixed effects of RI model with fixed hyperparamters: traceplot, autocorrelation and Geweke diagnostics.

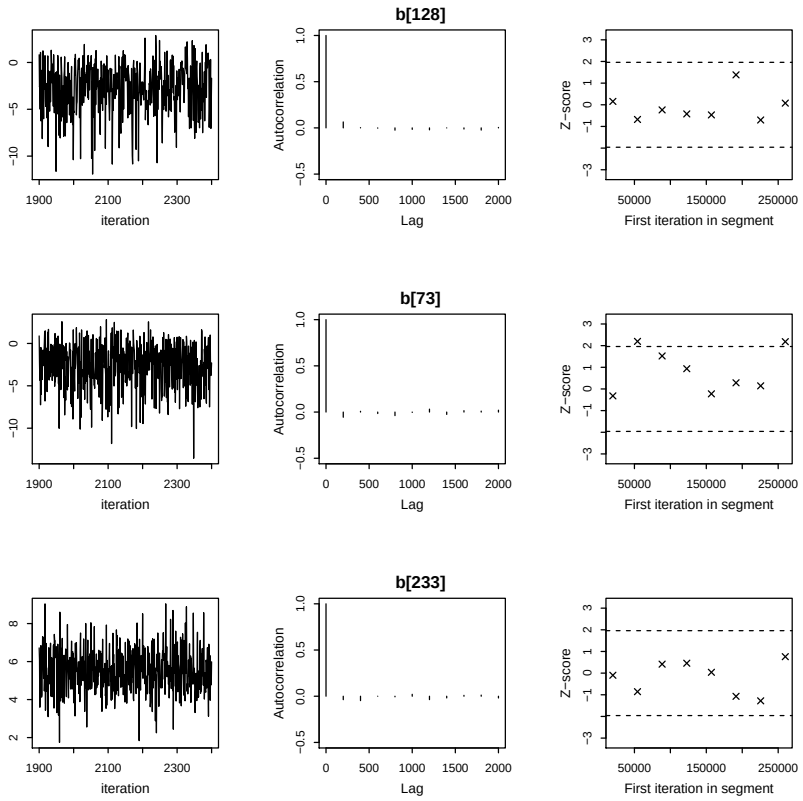

Figure 7: Convergence diagnostics for selected random effects of RI model with fixed hyperparameters: traceplot, autocorrelation and Geweke diagnostics.

2.3 Random intercept and slope model (RI+RS) with prior

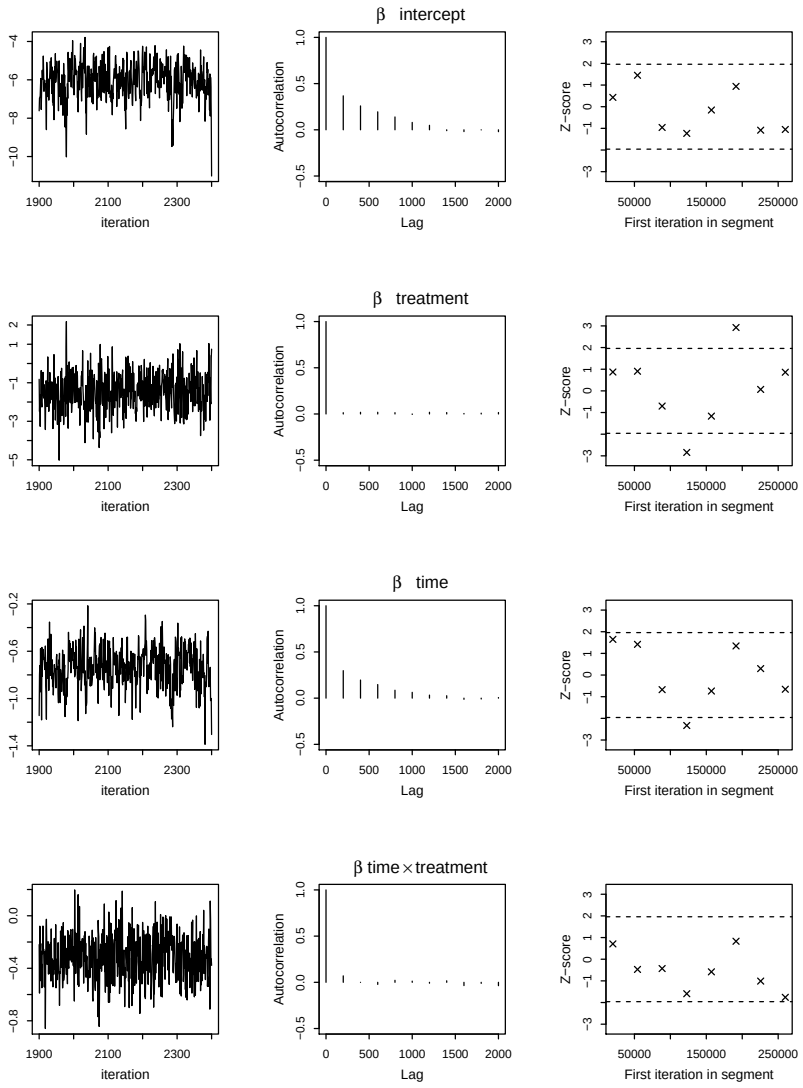

Figure 8: Convergence diagnostics for fixed effects of RI+RS model: traceplot, autocorrelation and Geweke diagnostics.

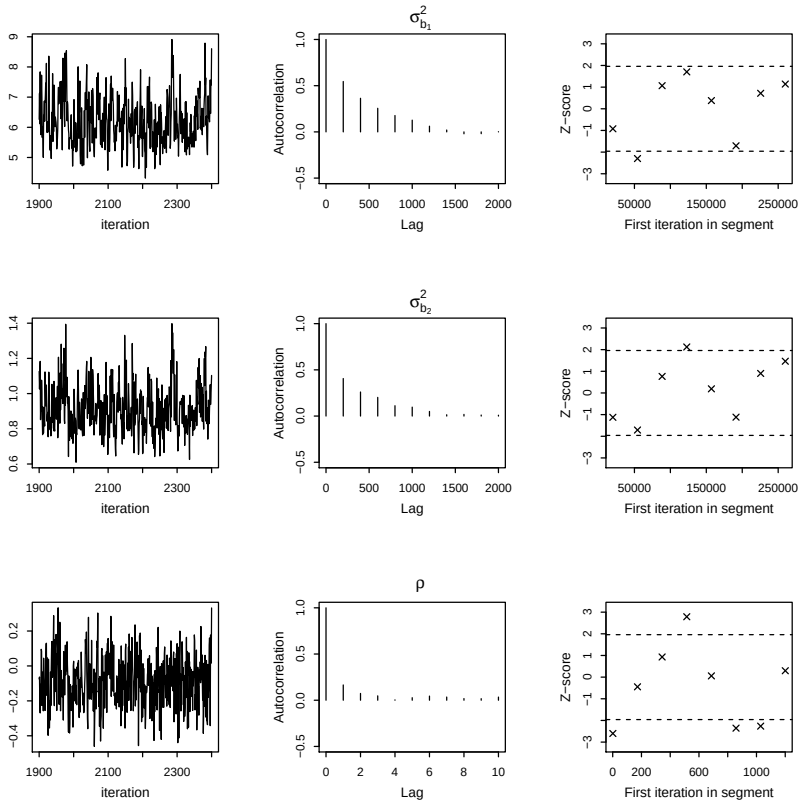

Figure 9: Convergence diagnostics for hyperparameters of RI+RS model: traceplot, autocorrelation and Geweke diagnostics.

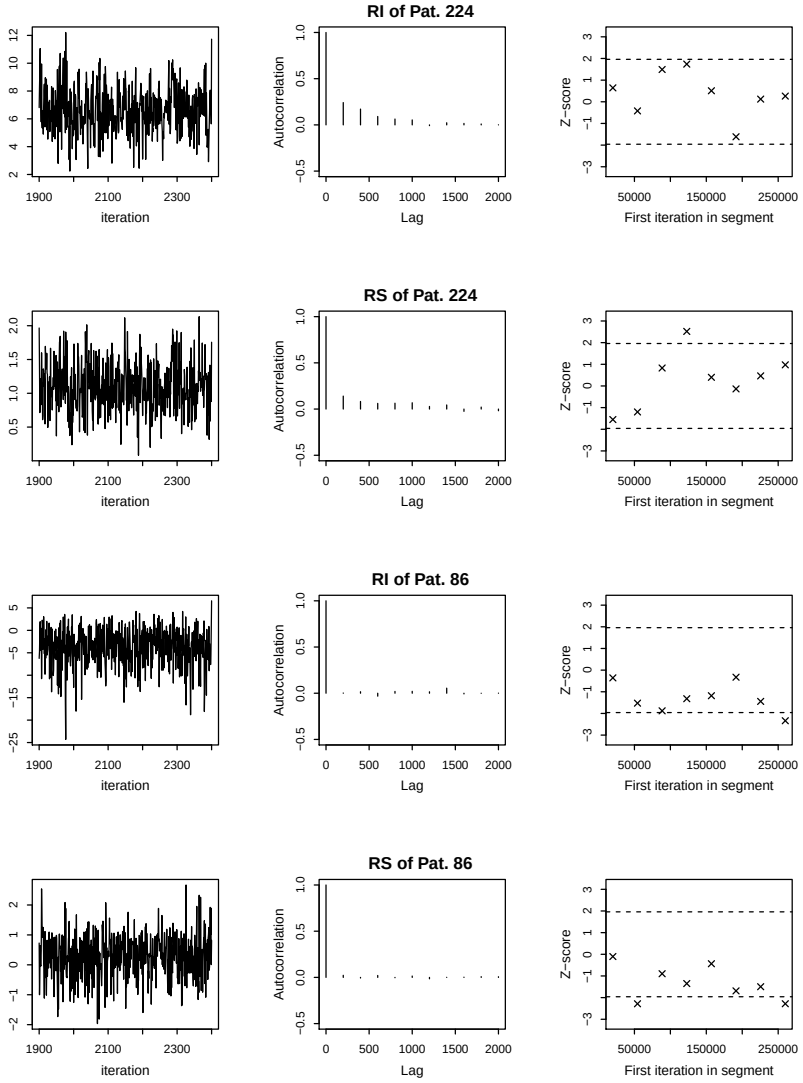

Figure 10: Convergence diagnostics for selected random effects of RI+RS model: traceplot, autocorrelation and Geweke diagnostics.

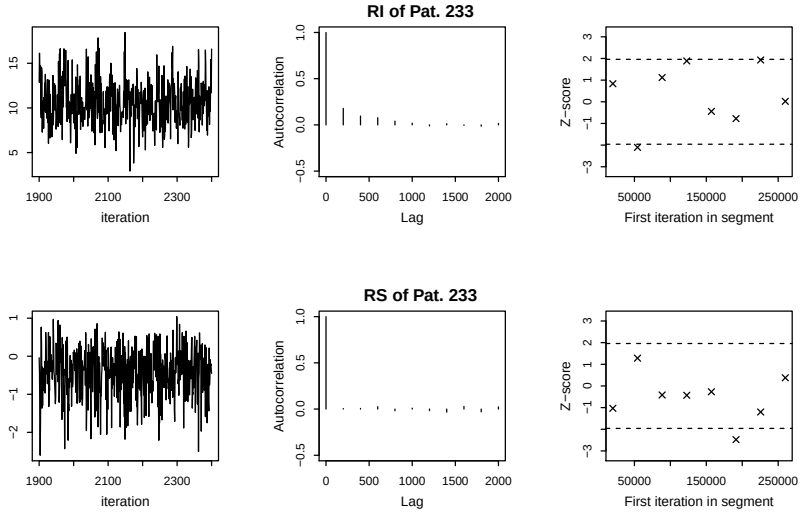

Figure 11: Convergence diagnostics for selected random effects of RI+RS model: traceplot, autocorrelation and Geweke diagnostics.

2.4 Random intercept and slope model (RI+RS) with fixed hyperparameters

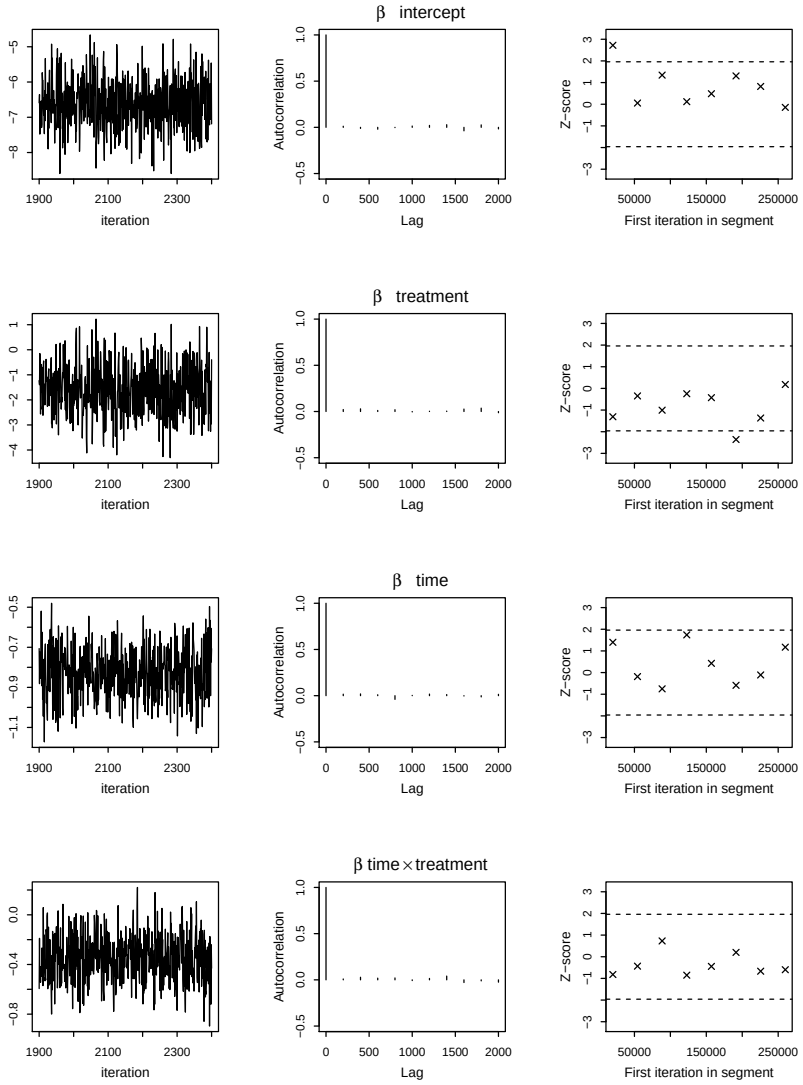

Figure 12: Convergence diagnostics for fixed effects of RI+RS model: traceplot, autocorrelation and Geweke diagnostics.

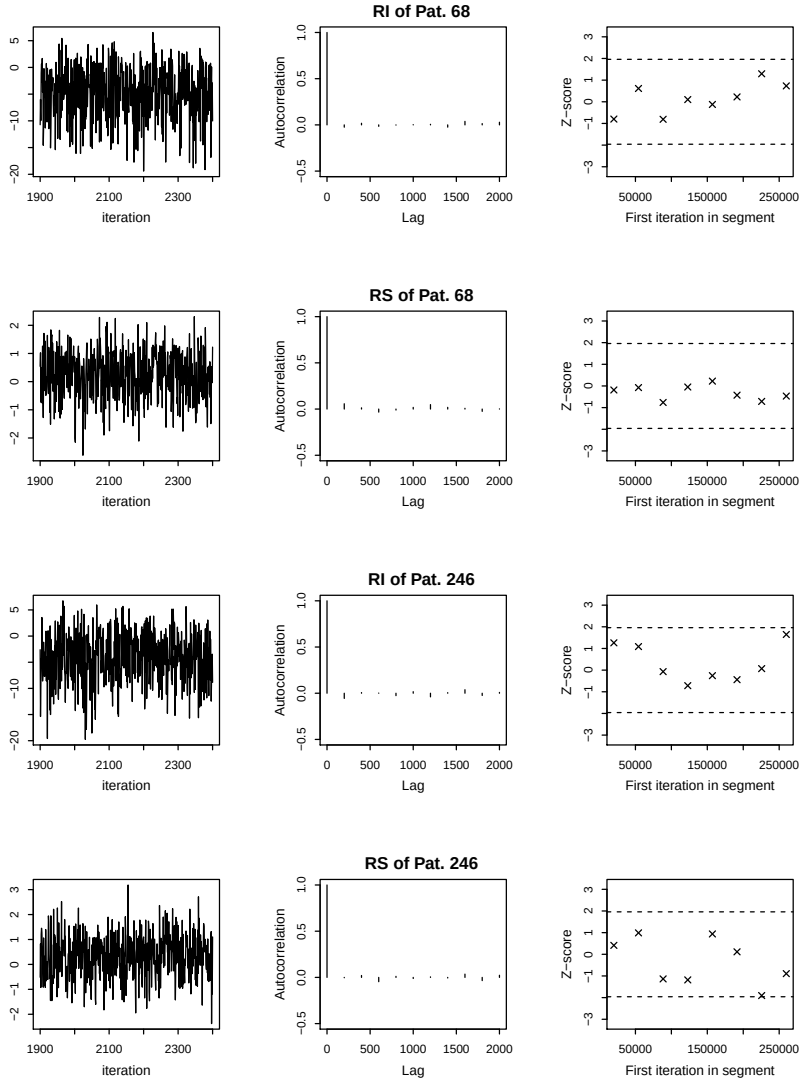

Figure 13: Convergence diagnostics for selected random effects of RI+RS model: traceplot, autocorrelation and Geweke diagnostics.

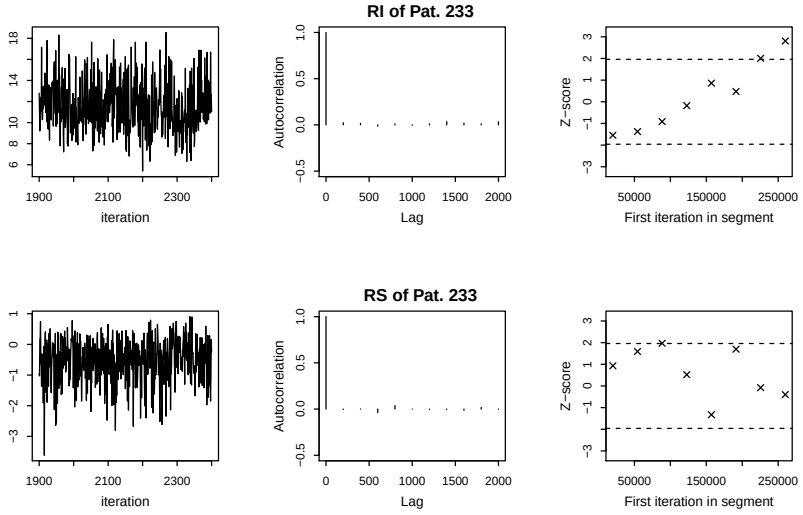

Figure 14: Convergence diagnostics for selected random effects of RI+RS model: traceplot, autocorrelation and Geweke diagnostics.

PAPER III

Network meta-analysis with integrated nested Laplace approximations

Rafael Sauter, Leonhard Held

Paper published in *Biometrical Journal*.

Network meta-analysis with integrated nested Laplace approximations

Rafael Sauter* and **Leonhard Held**

Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Hirschengraben 84, CH-8001 Zürich, Switzerland

Received 17 July 2014; revised 23 March 2015; accepted 27 May 2015

Analyzing the collected evidence of a systematic review in form of a network meta-analysis (NMA) enjoys increasing popularity and provides a valuable instrument for decision making. Bayesian inference of NMA models is often propagated, especially if correlated random effects for multiarm trials are included. The standard choice for Bayesian inference is Markov chain Monte Carlo (MCMC) sampling, which is computationally intensive. An alternative to MCMC sampling is the recently suggested approximate Bayesian method of integrated nested Laplace approximations (INLA) that dramatically saves computation time without any substantial loss in accuracy. We show how INLA apply to NMA models for summary level as well as trial-arm level data. Specifically, we outline the modeling of multiarm trials and inference for functional contrasts with INLA. We demonstrate how INLA facilitate the assessment of network inconsistency with node-splitting. Three applications illustrate the use of INLA for a NMA.

Keywords: Bayesian inference; Integrated nested Laplace approximations; Network meta-analysis; Node-splitting.

Additional supporting information including source code to reproduce the results may be found in the online version of this article at the publisher's web-site

1 Introduction

A systematic review on a particular outcome assembles the evidence of all available and relevant trials. A meta-analysis is the statistical tool used to analyze the results of the collected trials. The conventional meta-analysis performs pairwise treatment comparisons only. However, often one is not only interested in comparing two treatments but in a set of different interventions used to treat the same outcome. Instead of analyzing a series of pairwise comparisons one can describe the set of treatments as a network. The evidence, collected by the trials, is available for many but not necessarily for all possible direct pairwise treatment comparisons in this network. Still we can compare the remaining treatments indirectly as suggested by Bucher et al. (1997), relying on the evidence obtained by the observed direct comparisons. The conventional pairwise meta-analysis method in combination with the idea of indirect treatment comparisons is the cornerstone of a network meta-analysis (NMA) or mixed treatment comparison.

Analyzing a network of treatments enjoys increasing popularity resulting in an increasingly growing literature although this approach implies several challenges and is more complicated than the conventional pairwise meta-analysis. Mills et al. (2012) discuss possible biases caused by trials included in a NMA that are not sufficiently homogeneous or with too different interventions or study populations.

*Corresponding author: e-mail: rafael.sauter@uzh.ch

Mills et al. (2013) give an overview of the implications caused by between trial heterogeneity in a network, comparable to the heterogeneity in a pairwise meta-analysis. They also discuss the consequences of network inconsistencies. Results based on a NMA may easily be flawed if these aspects are not taken into account. See Salanti (2012) for a discussion about the concepts and assumptions behind a NMA and the growing importance of this method.

Generally, a NMA model can be estimated by maximum likelihood, but the hierarchical structure of the model may result in a complicated expression for the likelihood that requests numerical optimization. Alternatively Bayesian inference is often encouraged. The standard way for a Bayesian inference is Markov chain Monte Carlo (MCMC) sampling that is however computationally intensive. A fast and accurate alternative to MCMC has been proposed by Rue et al. (2009), the integrated nested Laplace approximations (INLA) of the marginal posterior distribution of the model parameters.

In this article, we discuss the inference of NMA with INLA and demonstrate that the estimates obtained by INLA are very close to the ones by MCMC. In Section 2, we first introduce two established NMA models, one for summary level and one for trial-arm level data and we show how they account for heterogeneity and inconsistency. In Section 3, we discuss Bayesian inference of NMA models with INLA. We place emphasis on the incorporation of multivariate random effects for multiarm trials, inference for functional contrasts and the implementation of the node-split approach to examine network inconsistencies (Dias et al., 2010). The implementation of NMA in INLA is further illustrated by three applications in Section 4. A brief discussion is provided in Section 5. Additionally, we provide R-code in the Supplementary Material that demonstrates the key implementation features of a NMA with INLA.

2 Statistical models for network meta-analysis

A network for a meta-analysis consists of a number of observed direct pairwise comparisons among T different treatments. The remaining relative effects among all possible pairwise comparisons are available by indirect comparisons, if a combination of direct treatment comparisons is observed that allows to form the indirect comparison. Different pairwise effect measures could be used in an NMA such as the odds ratio, the risk ratio or the risk difference. See Norton et al. (2012) and van Valkenhoef and Ades (2013) for a discussion about differences in NMA results if different effect measures are chosen. In this paper, we exclusively use the odds ratio as effect measure, but other measures could be used as well.

In Section 2.1, we introduce a NMA model for summary level data. Section 2.2 discusses a NMA model for trial-arm level data with a binomial outcome. In Section 2.2 the number of events and the total number of patients is reported on the basis of every trial-arm in the network, in contrast to summary level data that only report effect measures for each pairwise treatment comparison.

2.1 Summary level data

The model described in this section follows the one described by Lumley (2002). An effect measure y_{ijk}, here the log-odds ratio, comparing treatment j with k in trial i is observed together with its squared standard error σ^2_{ijk} for several independent two-arm trials $i = 1, 2, \ldots, S$. The treatment pair $k, j \in \{1, \ldots, T\}$ compared in trial i is one combination among $T(T-1)/2$ possible combinations. The log-odds ratio y_{ijk} is assumed to follow a normal distribution and is modeled as

$$y_{ijk} \sim \mathrm{N}\big(d_{jk} + \gamma_{ijk}, \sigma^2_{ijk}\big). \tag{1}$$

The relative treatment effect d_{jk} is the difference between the treatment effects d_j and d_k, say, such that $d_{jk} = d_j - d_k$. The observed squared standard error σ^2_{ijk} of each effect measure is used as an inverse

weight to scale the model variance. The model allows for an additional source of uncertainty by introducing variation through random effects γ_{ijk}. Excessive variation between trials, called heterogeneity, is captured by the random effects $\gamma_{ijk} \sim \mathrm{N}(0, \tau^2)$. This trial-specific heterogeneity captures differences between trials comparing the same treatments but being different in terms of trial-specific features, for example differences between study-populations. The random effects variance τ^2 is a measure for the degree of heterogeneity in the network. A large random effect γ_{ijk} indicates that there is a between-trial variability exceeding the expected sampling variability for the treatment comparison j versus k in trial i.

The model for summary level data contains only information about pairwise treatment differences and includes thus two-arm trials only. In order to make the model identifiable, we need to fix the treatment effect of some arbitrary baseline treatment at zero. By consequence, we only need $T - 1$ parameters to fully describe the model with its network structure. If we choose treatment 1 as baseline then we define the *basic contrasts* $\boldsymbol{d}_b = (d_{12}, d_{13}, \ldots, d_{1T})^\top$, the treatment effects relative to the baseline treatment. Based on $\boldsymbol{d}_b$ we can fully describe the network structure. The relative treatment effects not contained in $\boldsymbol{d}_b$, the so-called *functional contrasts* $\boldsymbol{d}_f$, can be expressed as linear combinations of $\boldsymbol{d}_b$ under the assumption of consistency. Network consistency means that there is no discrepancy between the evidence obtained from direct and indirect comparisons. To illustrate this, assume that we have three treatments 1, 2, 3 with all three possible pairwise treatment comparisons being observed. If treatment 1 is chosen as baseline then the basic contrasts are $\boldsymbol{d}_b = (d_{12}, d_{13})^\top$, but we may also be interested in the functional contrast $d_f = d_{23}$. Under consistency, we have $d_{23} = d_{13} - d_{12}$, i. e. $d_f = \mathbf{F}^\top \boldsymbol{d}_b$ where $\mathbf{F} = (1, -1)^\top$. If the network consists of N pairwise comparisons, then the number of functional contrasts is $N - T + 1$.

A crucial component of every NMA is the assessment of network inconsistencies, that is the examination of the possibility that consistency restrictions are not fulfilled. If the equality of indirect and direct comparisons does not hold, it is possible to capture this inconsistency by introducing additional random effects $\xi_{jk} \sim \mathrm{N}(0, \kappa^2)$ in Eq. (1),

$$y_{ijk} \sim \mathrm{N}(d_{jk} + \gamma_{ijk} + \xi_{jk}, \sigma^2_{ijk}), \tag{2}$$

where the variance κ^2 is a measure for the degree of inconsistency in the network. Consistency of a pairwise comparison between treatments j and k is thus put into doubt if we have a large random effect estimate of ξ_{jk}.

2.2 Trial-arm level data

The model discussed in this section has been introduced by Lu and Ades (2006). Compared to the model for summary level data of Section 2.1, where every trial is assumed to have two arms only, it is possible in a model for trial-arm level data to account for multiarm trials. Each trial $i = 1, 2, \ldots, S$ has treatment arms $t_1(i), \ldots, t_{K_i}(i) \in \{1, \ldots, T\}$ with at least $K_i \geq 2$ treatment arms. The first treatment $j = t_1(i)$ is chosen as baseline treatment and compared with the remaining treatments $k = t_2(i), \ldots, t_{K_i}(i)$. For each trial i and baseline treatment j the number of events y_{ij} and number of patients n_{ij} is observed. Correspondingly, also for the remaining treatments y_{ik} and n_{ik} is observed. The number of events is (conditionally) independent for each trial-arm and follows a binomial distribution, that is $y_{ij} \sim \mathrm{Bin}(n_{ij}, \pi_{ij})$ as well as $y_{ik} \sim \mathrm{Bin}(n_{ik}, \pi_{ik})$. The log-odds ratio d_{jk} of baseline treatment j versus treatment k can now be modeled with logistic regression as

$$\mathrm{logit}(\pi_{ij}) = a_{ij} \tag{3}$$

$$\mathrm{logit}(\pi_{ik}) = a_{ij} + d_{jk} + \gamma_{ijk}. \tag{4}$$

The treatment effect a_{ij} of baseline treatment j in trial i is a nuisance parameter and the main interest is in the log-odds ratio d_{jk}.

Similar to the model for summary level data, possible trial-specific heterogeneity is captured by the random effects $\gamma_{ijk} \sim \mathrm{N}(0, \tau^2)$. However, treatment comparisons in a multiarm trial with more than two treatments are not independent, because they are based on the same baseline data. For example if trial i compares treatments 1, 2, and 3, we take the dependency into account by assuming a multivariate normal distribution for the random effects vector $\boldsymbol{\gamma}_i = (\gamma_{i12}, \gamma_{i13})^\top$. In general, in a multiarm trial i with K_i different treatments, $\boldsymbol{\gamma}_i$ is a vector of length $(K_i - 1)$ and follows a multivariate normal distribution

$$\boldsymbol{\gamma}_i \sim \mathrm{N}(\mathbf{0}, T_i)$$

where T_i is a symmetric covariance matrix of dimension $(K_i - 1) \times (K_i - 1)$. A fully parametrized, unstructured covariance matrix with $(K_i - 1)(K_i - 2)/2$ parameters is not very practical as usually there are not many multiarm trials comparing the same set of treatments in a network. Therefore a reduction of the number of parameters in order to increase the efficiency of the parameter estimates is often warranted. Higgins and Whitehead (1996) suggest an exchangeable or homogeneous covariance matrix where all treatment-specific variances on the diagonal of T_i are set to τ^2. If we assume consistency in the network then the correlations between random effects for any two treatments of the same trial are equal to $\rho = 1/2$ (see Higgins and Whitehead, 1996, Section 5.1). This implies that covariances in T_i are under consistency all equal to $\tau^2/2$. Throughout the remainder we use such a homogeneous correlation matrix for T_i.

Basic and functional contrasts and modelling of network inconsistency through random effects is different for models based on trial-arm level data compared to models for summary level data described in Section 2.1. The vector of basic contrasts $\boldsymbol{d}_b$ of length $T - 1$ in a NMA for trial-arm level data can now be any set of directly observed effect parameters, which define a spanning tree (see Lu and Ades, 2006, Section 2.3), that is a connected sub-graph of the network covering all vertices without any loops. The remaining functional contrasts $\boldsymbol{d}_f$ can again be described as linear combinations of $\boldsymbol{d}_b$.

As the equality of indirect and direct comparisons in a network may not be fulfilled, we introduce additional random effects that allow for deviations from the consistency restrictions. The loop-specific approach by Lu and Ades (2006) proposes that for every independent three-way loop a random effect is added (see also Dias et al., 2010). In contrast to the situation with pairwise comparisons, the presence of multiarm trials imply that the number of random effects is not necessarily the same as the number of functional contrasts. However, if a direct comparison is only observed in one multiarm trial, then any indirect comparison based on the other treatments in the same trial does not form an independent loop. Thus, multiarm trials are assumed to be inherently consistent. Consider a three-way loop with treatments j, k, l, where we now introduce a random effect $\xi_{jkl} \sim \mathrm{N}(0, \kappa^2)$ in model (4). To do so, we relax the consistency relation $d_{lk} - d_{lj} = d_{jk}$ to $d_{lk} - d_{lj} = d_{jk} + \xi_{jkl}$ and hence replace d_{jk} with $d_{jk} + \xi_{jkl}$ in (4):

$$\mathrm{logit}(\pi_{ik}) = a_{ij} + d_{jk} + \gamma_{ijk} + \xi_{jkl}. \tag{5}$$

As before, the degree of inconsistency is represented by the random effects variance κ^2 and a large estimate of ξ_{jkl} indicates a discrepancy between the direct and indirect evidence. Identification of all independent three-way loops can be complicated for a network with many multiarm trials. Furthermore, parametrization of the inconsistency random effects is not unique in combination with correlated multivariate heterogeneity random effects (see Sections 4.2 and 4.5 in Lu and Ades, 2006). An alternative to infer inconsistency by including additional random effects is offered by node-splitting, see Section 3.3.

3 Bayesian inference for network meta-analysis models with integrated nested Laplace approximations

The NMA models discussed above are hierarchical models that can be described by three stages: the first stage is the observational model $p(\mathbf{y} \mid \boldsymbol{\alpha})$ with respect to the observed data $\mathbf{y}$ where $\boldsymbol{\alpha} = (\boldsymbol{a}, \boldsymbol{d}_b, \boldsymbol{\gamma}, \boldsymbol{\xi})$ includes all model parameters. Here $\boldsymbol{a} = (a_{1j}, \ldots, a_{Sj})^\top$ denotes the vector of all baseline treatment effects in the model for trial-arm level data and is not needed for a model based on summary level data. The random effects vector $\boldsymbol{\gamma}$ contains all trial-specific random effects $\boldsymbol{\gamma}_i$. Likewise, $\boldsymbol{\xi}$ contains all incoherence random effects ξ_{jk} of the model based on summary level data or ξ_{jkl} in the case of trial-arm level data. The second stage $p(\boldsymbol{\alpha} \mid \boldsymbol{\theta})$ is a latent Gaussian Markov random field (GMRF) as described in chapter 2 of Rue and Held (2005). The GMRF is controlled by hyperparameters, here $\boldsymbol{\theta} = (\tau^2, \kappa^2)^\top$, which build the third stage $p(\boldsymbol{\theta})$.

The INLA approach proposed by Rue et al. (2009) generates accurate approximations to the marginal posterior distributions of the latent Gaussian model $\boldsymbol{\alpha}$ by applying a Laplace approximation (Tierney and Kadane, 1986) to the posterior distribution of $\boldsymbol{\theta}$ and a second Laplace approximation to the posterior of the components of $\boldsymbol{\alpha}$ for selected values of the hyperparameters. The approximation of the marginal posterior distributions to the model parameters $\boldsymbol{\alpha}$ is obtained by numerical integration over the hyperparameters. INLA were shown to deliver accurate approximations with reduced computational costs compared to MCMC in a variety of examples. Fong et al. (2010) demonstrated the applicability of INLA to generalized linear-mixed models. See Schrödle et al. (2011) for applications to spatio-temporal models and Paul et al. (2010) for a bivariate meta-analysis of diagnostic tests with INLA. The available software package for INLA embraces a wide range of models that is progressively expanded as for example by measurement error models described by Muff et al. (2015).

The software `r-inla` implements the INLA approach, is available on `http://www.r-inla.org/` and includes an `R`-interface. In the following Sections 3.1 and 3.2, we discuss the implementation of NMA in `r-inla` for summary and a trial-arm level data models. The node-split approach with `r-inla` is introduced in Section 3.3. Each feature discussed in Subsections 3.1 to 3.3 is complemented by an application in Section 4. We compare the results obtained with the INLA approach with MCMC, where we rely on the `R`-package `gemtc` (van Valkenhoef and Kuiper, 2014) using `JAGS` (Plummer, 2003) a generic MCMC sampler using BUGS-code. However, the `gemtc`-package can not fit models with both heterogeneity and incoherence random effects, as shown in Table 1 that gives an overview of which models can be estimated with the utilized software packages. The NMA model with incoherence, for summary level data could be estimated by extending the `gemtc`-code and for trial-arm level data the BUGS-code provided by Lu and Ades (2006) on `http://www.bristol.ac.uk/social-community-medicine/projects/mpes/code/` was adapted to fit into `JAGS`.

Table 1 Used software packages for NMA: '`x`' can be estimated, '`na`' not available.

	`r-inla`	`gemtc`	`JAGS`
NMA models for summary level data:			
random effects (heterogeneity + inconsistency)	x	na	x
NMA models for trial-arm level data:			
fixed effects	x	x	x
random effects (heterogeneity)	x	x	x
random effects (heterogeneity + inconsistency)	x	na	x

Source code for MCMC sampling and the `r-inla` implementation and datasets together with an accompanying `R`-package to reproduce the results is available as Supporting Information on the journal's web page.

3.1. Summary level data

The NMA model for summary level data described in Eq. (2) is a linear mixed model. Linear-mixed models with different random effects can be implemented in `r-inla` (Fong et al., 2010) and Schmidt and Nehmiz (2014) used INLA to estimate such a NMA model with hazard ratios as effect measures and compared the results to the ones by MCMC.

However, there are two peculiarities in the linear-mixed model (2) that need special attention: weighting the outcome with the inverse of the squared standard error σ^2_{ijk} and inference for the functional contrasts $\boldsymbol{d}_f$. As in a standard meta-analysis, we want to use σ^2_{ijk} as inverse weight. This can be done in `r-inla` by using the `scale` argument in the function call. As discussed in Sections 2.2 and 2.1, the functional contrasts $\boldsymbol{d}_f$ can be described as linear combination of $\boldsymbol{d}_b$ and we want to get the marginal posterior distributions of $\boldsymbol{d}_f$. The `r-inla` software package allows to compute the marginal posterior distribution of any linear combination of the latent field (see Section 4.4 in Martins et al., 2013) and thus the full marginal distributions for $\boldsymbol{d}_f$ can be obtained. If one is interested in the functional contrasts one has to define these linear combinations in advance and hand them over in the `lincomb` argument of the `r-inla` function call. Further implementation details for these two peculiarities are described in the Supplementary Material.

3.2 Trial-arm level data

The NMA model for trial-arm level data described by (3) to (5) is a binary regression model with logit link function. The distribution of the response must be specified in `r-inla` by the `family` argument and the total number of study participants needs to be specified by the argument `Ntrials`. Functional contrasts can be obtained in `r-inla` by building linear combinations in the same way as described in Section 3.1. In the case of a multiarm trial we want to correlate the random effects of the latent field for all arms of the same trial i, as discussed in Section 2.2. The use of such correlated multivariate random effects is possible in `r-inla`. Riebler et al. (2012) use INLA for correlated multivariate age-period-cohort models. Their model includes several countries and for each country a second order random walk over time is used. These random walks are correlated across countries. Here, we want to introduce an exchangeable correlation structure across trial-arms. Correlating the random effects for the outcomes of a three-arm trial means that we correlate two components of the latent field.

To implement the correlation across study arms in `r-inla` we need to rewrite the homogeneous random effects covariance T_i. Using the uniform correlation matrix $\mathrm{C}_i = (1-\rho)\mathrm{I}_i + \rho\mathrm{J}_i$, where I_i is the identity matrix and J_i is a matrix of ones, both of dimension $(K_i - 1)$, we can write $T_i = \mathrm{C}_i \otimes \tau^2$ where $\otimes$ is the Kronecker product. This describes the random effects correlation structure for the response vector $\boldsymbol{y}_i$ of trial i. As described in Section 2.2, we need to fix the hyperparameter ρ to the initial value $1/2$ but this initial value must be transformed to the internal scale used by `r-inla` (see Section 1.2 in Supplementary Material). The homogeneous correlation model described here is implemented in `r-inla` under the name `model='exchangeable'`.

To identify the corresponding random effects in each trial, we need to define a grouping vector $\mathbf{g}_i$ that is needed to determine the structure of C_i. The first entry of $\mathbf{g}_i$ is `NA` as no random effect is present in the baseline model (3). The remaining entries are numbered from 1 to $K_i - 1$, representing the random effects $\gamma_{ij1}, \ldots, \gamma_{ij,(K_i-1)}$ in model (4). For example, if trial i compares treatments 1, 2, and 3, we have a bivariate random effects vector $\boldsymbol{\gamma}_i = (\gamma_{i12}, \gamma_{i13})^\top$ and the grouping vector is $\mathbf{g}_i = (\mathtt{NA}, 1, 2)^\top$. See section 4.6 in Martins et al. (2013) and the Supplementary Material for further details about this `r-inla` feature.

3.3. Assessing inconsistency with node-splitting

A key element of every NMA is to assess the degree and source of inconsistency in the network. Evidence inconsistencies can be modeled by random effects for treatment pairs (Section 2.1) or for independent loops (Section 2.2). An alternative to the random effects approach is to compare only the direct evidence of a relative treatment effect with the indirect evidence for the same relative treatment effect. This procedure corresponds to a cross-validation of the network leaving out all observations covering direct evidence on a specific treatment pair, as discussed by Lumley (2002, Section 5.2) for summary level, and by Lu and Ades (2006, Section 5.3) for trial-arm level data.

A modification of cross-validation is node-splitting as suggested by Dias et al. (2010, Section 3.5). As in cross-validation, the data are split into two independent sources of information on direct and indirect evidence for a pair of treatments j and k, say. However, in the node-split approach the heterogeneity hyperparameter τ^2 is estimated based on all the data. Assessing inconsistency by node-splitting implies that the model is fitted repeatedly for every directly observed comparison of treatments j and k where also indirect evidence is available. For every model fit, we obtain a posterior distribution for the hyperparameter τ^2, for the baseline treatment effects $\boldsymbol{a}$ and for the direct and indirect relative treatment effects $d_{jk}^{\text{dir.}}$ and $d_{jk}^{\text{ind.}}$.

The node-split approach can be implemented in `r-inla` using a joint model with two separate likelihood functions. Specifically, suppose treatment j and k are directly compared in s trials $D_{jk} \subset \{1, \ldots, S\}$. In trial $i \in D_{jk}$ there are 2 responses, y_{ij} and y_{ik}, which form the direct evidence and are incorporated as vector

$$\mathbf{Y}_{(j,k)} = (y_{i_1 j}, y_{i_1 k}, y_{i_2 j}, y_{i_2 k}, \ldots, y_{i_s j}, y_{i_s k})^\top$$

in the first likelihood. There are $K = \sum_{i=1}^{S} K_i$ treatment arms in total, so the indirect evidence is represented by the remaining $K - 2s$ responses, incorporated as vector $\mathbf{Y}_{-(j,k)}$, say, in the second likelihood. The unknown parameters $\boldsymbol{a}$, $\boldsymbol{d}_b$, and $\boldsymbol{\gamma}$ must be organized accordingly in the two likelihoods. Of note, `r-inla` allows that the two likelihoods share some common parameters, here the heterogeneity variance τ^2. We can therefore use `r-inla` to compute the posterior distribution of the difference of the direct treatment effect $d_{jk}^{\text{dir.}}$ and the indirect treatment effect $d_{jk}^{\text{ind.}}$ from the first and second likelihood, respectively, as a linear combination.

Martino et al. (2011) describe a joint model where one likelihood defines data for survival times while the second likelihood defines a quantitative measure for a longitudinal profile of the same patient. They implement a joint model in `r-inla` with two likelihoods where a hyperparameter for the frailties in the survival likelihood and the random effects in the longitudinal likelihood is shared. See Martins et al. (2013, Section 4.1) for further details on how to use two likelihoods in `r-inla`.

4 Applications

In this section, we present three different applications: a NMA based on summary level data in Section 4.1, one based on trial-arm level data in Section 4.2 and node-splitting for a NMA in Section 4.3. We assume the same prior distributions for the components of $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ as in Lu and Ades (2006) and Dias et al. (2010). Specifically, for all components of $\boldsymbol{\alpha}$ we assume independent normal priors with mean zero and variance 1000. Independent uniform priors on the interval [0, 10] are used for the random effects standard deviations τ and κ.

The implementation of the user-specified uniform prior on the hyperparameters, which is achieved in `r-inla` by handing over a table with the prior density evaluated at an appropriate grid, is discussed in detail in the Supplementary Material. As in Lu and Ades (2006), 20,000 iterations with an additional burnin of 30,000 iterations are used for MCMC analyses in Sections 4.1 and 4.2. In Section 4.3, we used 300,000 iterations and an additional burnin of 200,000 samples to get the posterior distributions

for each node-split. The number of iterations is the same as used by Dias et al. (2010) who mention that most of the models reach convergence for much less iterations while some of the node-splits need many sampling iterations to satisfy diagnostic convergence criteria.

4.1 Acute myocardial infarction summary level data

An acute myocardial infarction is caused by a clot in a coronary artery. There exist different interventions to remove the clot that have been compared by several trials. The summary level data discussed here are taken from Lumley (2002, Table 2 in Section 5.2.). The data describe a network of six different treatments after a myocardial infarction (1: streptokinase, 2: t-PA, 3: accelerated t-PA, 4: reteplase, 5: anistreptilase, 6: angioplasty). There are 16 directly observed treatment comparisons covering nine different among 15 possible treatment pairs for which the log-odds ratio and the corresponding squared standard error σ^2_{ijk} are reported. We choose streptokinase (treatment 1) as baseline by setting its coefficient to zero. Thus $\boldsymbol{d}_b$ contains the five relative treatment effects to the baseline treatment 1. Figure 1 shows the marginal posterior densities for all basic contrasts and for one functional contrast d_{43} (reteplase vs. accelerated t-PA). The last two plots in Fig. 1 show the marginal posterior densities for the two hyperparameters τ^2 and κ^2 that are both very close to zero suggesting that there is neither a large heterogeneity nor a large incoherence in the network.

The histograms illustrate the densities based on the MCMC samples and the straight lines show the corresponding densities obtained by `r-inla`. Both methods MCMC and INLA produce very similar results. The largest absolute difference for the posterior median estimate of the log-odds ratio relative treatment based on MCMC and INLA among all basic contrasts was found for the contrast d_{16} (0.015). The MCMC run with 50,000 iterations took approximately 2.1 seconds while `r-inla` only took 1.1 seconds.

4.2 Smoking cessation trial-arm level data

The effect of four interventions, which support participants in their efforts to quit smoking, were compared by several trials originally discussed by Hasselblad (1998) but also by Lu and Ades (2006), Dias et al. (2010), and Kessels et al. (2013). The trial-arm level data measures the number of individuals who successfully quit smoking after 6–12 months. Data are taken from Table 1 in Lu and Ades (2006). The smoking cessation data describes a fully connected network comparing the effects of four different interventions (1: self-help, 2: individual counselling, 3: group counselling, and 4: no contact) and reports the number of successes and number of participants in 24 trials. There are two three-arm trials, one for treatments 1, 3, and 4 and one for treatments 2, 3, and 4. As baseline treatment we choose intervention 1 such that the basic contrasts $\boldsymbol{d}_b = (d_{12}, d_{13}, d_{14})^\top$ define a spanning tree of the graph. As in Lu and Ades (2006) we analyze three models, one with fixed effects setting $\tau^2 = 0$, one with random effects for heterogeneity and a model with random effects for heterogeneity and loop-specific inconsistency. The inconsistency in the third model is described by the three random effects $\xi_{123}, \xi_{124}, \xi_{134}$, one for each functional contrast $d_{23}, d_{24}, d_{34,}$ and each specific for one of the three independent loops in the network.

Figure 2 shows the posterior median and the 95% equi-tailed credible interval (CI) obtained by INLA and by MCMC for all parameter estimates of the three models. The results in Fig. 2 are consistent with the results reported by Lu and Ades (2006) in Table 2. Due to illustrative purposes, we present the hyperparameters in Fig. 2 as standard deviations τ, κ, instead of variances. The median and the 95%-CI for τ in Fig. 2 show that there is substantial heterogeneity present in the network. The inconsistency random effects standard deviation κ with the 2.5% quantile very close to zero is lower than the standard deviation for heterogeneity τ. Correspondingly, the three random effect estimates $\xi_{123}, \xi_{124}, \xi_{134}$ are close to zero. Adding the inconsistency parameters, additional to heterogeneity

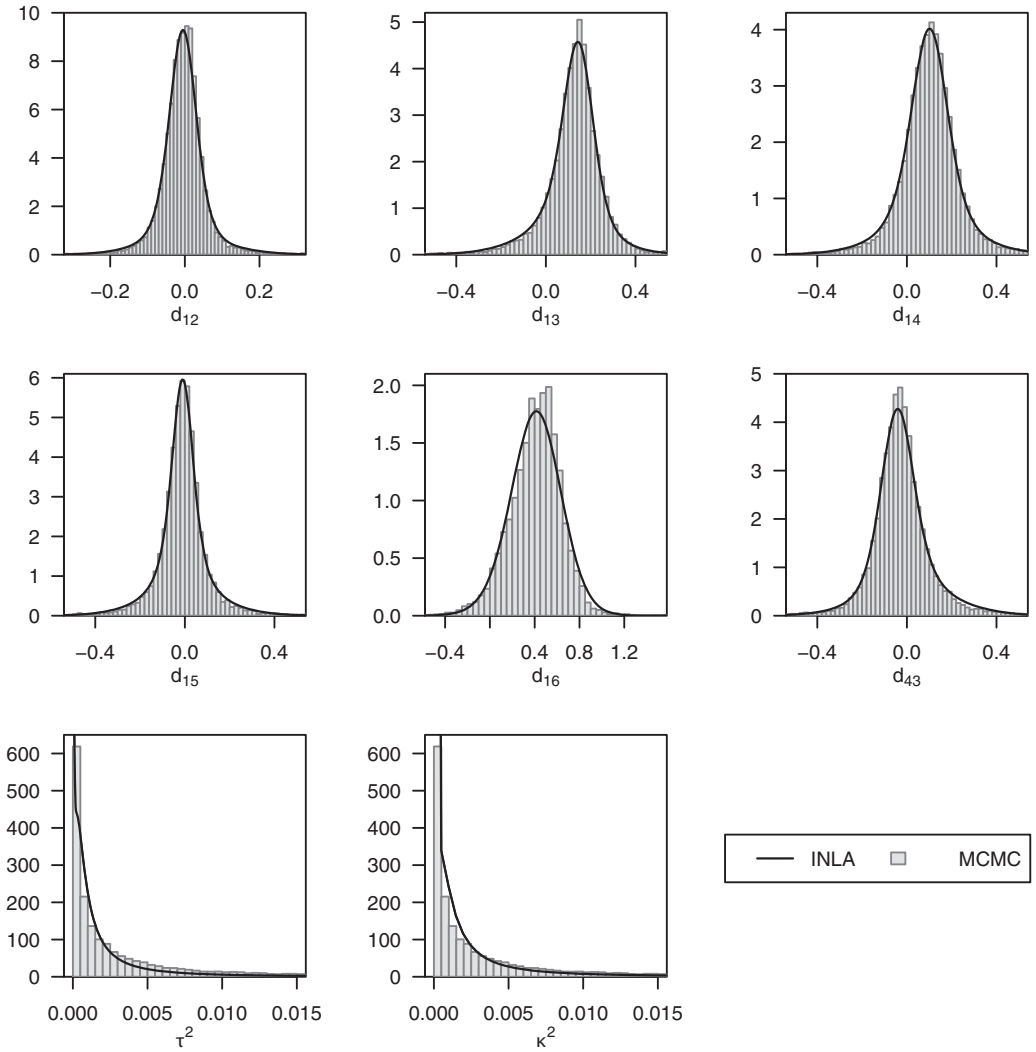

Figure 1 Marginal posterior density estimates of all basic contrasts relative to baseline treatment, for one functional contrast d_{43} and for the hyperparameters τ^2, κ^2 by MCMC (histogram) and by INLA (straight line) for the myocardial infarction data.

random effects, does not have a large impact on the estimates for the contrasts $\boldsymbol{d}_b$ and $\boldsymbol{d}_f$. A table with the numbers illustrated in Fig. 2 is available in the Supplementary Material.

The posterior distributions by MCMC and INLA for the basic and functional contrasts show all very good agreement. The largest absolute difference between the posterior median of both methods among all comparisons in the fixed effects model is for d_{12} but is still very small (0.0031). Similarly, the largest difference in the model with heterogeneity random effects is found for d_{13} with (0.0047). The model with additional inconsistency random effects shows the largest absolute difference for the contrast d_{23} (0.0125). Of course the differences and the ordering of the relative treatment comparisons

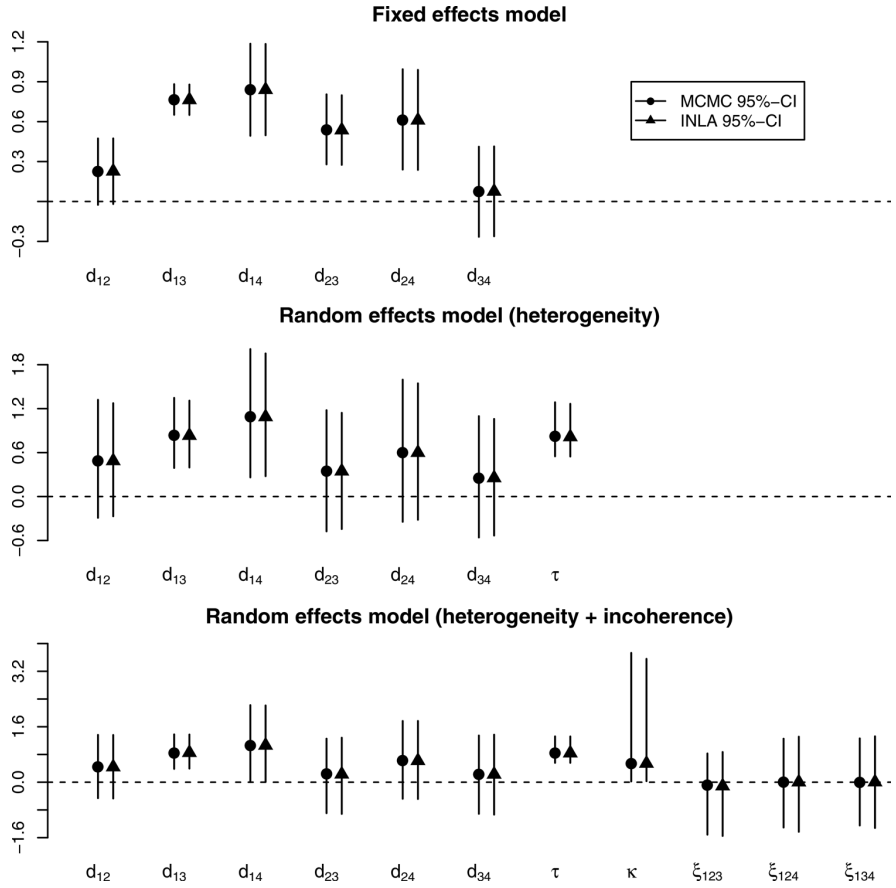

Figure 2 Median and 95% equi-tailed credible interval (CI) of the marginal posterior distributions of all relative treatment effects, the heterogeneity τ, and incoherence κ standard deviation as well as the incoherence random effects ξ_{1kc} by MCMC (points) and by INLA (triangles) for the smoking cessation data.

with the largest discrepancies between MCMC and INLA may change if the MCMC run is repeated while the results by INLA are deterministic. The MCMC run with 50,000 iterations took approximately 10.8 seconds for the model with random effects for heterogeneity and incoherence. With `r-inla` the computing time was 1.6 seconds.

4.3 Node-splitting for thrombolytic drugs

This application, like the one in Section 4.1, compares treatments after an acute myocardial infarction. Here, treatments are limited to thrombolytic drugs, whereas in Section 4.1 also physical interventions by angioplasty were included in the analysis and data are on trial-arm instead of summary level. Data are taken from Dias et al. (2010) but are also discussed in Lu and Ades (2006) in Table 3. The dataset compares nine different treatments and reports the number of deaths in 30 or 35 days and number of patients in each treatment arm for 50 different trials. There is direct evidence for 16 different pairwise

Table 2 Posterior mean and standard deviation of the difference $d_{jk}^{\text{diff.}}$ of direct and indirect log odds ratios with corresponding conflict p-value for all node-splits of the thrombolytic treatment network. All results have been obtained with INLA. Left column is based on the fixed effects model, right column based on the random effects model.

$d_{jk}^{\text{diff.}}$	Fixed effects model			Random effects model		
j, k	Mean	Std. dev.	p-value	Mean	Std. dev.	p-value
1, 2	−0.19	0.23	0.43	−0.24	0.28	0.38
1, 3	0.09	0.10	0.39	0.25	0.25	0.28
1, 5	0.12	0.12	0.34	0.40	0.34	0.20
1, 7	−0.27	0.22	0.22	−0.23	0.25	0.35
1, 8	−0.18	0.56	0.75	−0.13	0.58	0.83
1, 9	−0.41	0.25	0.10	−0.39	0.28	0.16
2, 7	−0.05	0.42	0.92	0.02	0.45	0.97
2, 8	−0.14	0.45	0.75	−0.13	0.47	0.79
2, 9	−0.43	0.24	0.077	−0.51	0.29	0.071
3, 4	−0.65	0.67	0.33	−0.87	0.73	0.22
3, 5	−0.12	0.12	0.33	−0.40	0.34	0.19
3, 7	0.26	0.21	0.22	0.21	0.24	0.37
3, 8	0.27	0.45	0.56	0.22	0.48	0.66
3, 9	1.20	0.41	0.001	1.21	0.43	0.002

treatment comparisons but there are only 14 independent loops for which a node-split is possible. We performed the node-splitting of a fixed effect model, setting τ^2 to zero and a model including heterogeneity random effects. We use the same measure of inconsistency as proposed by Dias et al. (2010). They define the measure for the degree of inconsistency ($d_{jk}^{\text{diff.}}$) as the difference of the log-odds ratios based on direct and indirect evidence, that is $d_{jk}^{\text{diff.}} = d_{jk}^{\text{dir.}} - d_{jk}^{\text{ind.}}$. We thus compute the posterior distribution of $d_{jk}^{\text{diff.}}$ from which a two-sided conflict p-value (Marshall and Spiegelhalter, 2007) can be easily derived.

The inconsistency estimates obtained by INLA are shown in Table 2, both for the fixed and the random effects model. The fixed effects estimates are very similar to the MCMC estimates shown in (Dias et al., 2010, Table 2). We also used MCMC sampling for the node-splitting of the same network with both the fixed and random effects models. Inconsistency estimates by INLA agreed well with the results by MCMC. Detailed results can be found in Section 2 in the Supplementary Material. Of note, the node-split analysis of the moderately large thrombolytic treatment network with 14 node-splits and heterogeneity random effects took 3.2 hours with MCMC, approximately 14 minutes per node-split. The corresponding NMA model has 60 parameters and hyperparameters, so carrying out MCMC sampling for all 14 node-split analyses means that we need to check convergence of 840 parameters. `r-inla` was 276 times faster and needed only 42 seconds to complete the 14 node-splits that is 3 seconds per node-split.

5 Discussion

The application of INLA to NMA models in Sections 4.1, 4.2, and 4.3 showed results very close to those obtained with MCMC. However, computation time with INLA is drastically reduced compared

to MCMC sampling. Further, there is no need to examine convergence of the MCMC samples. This is a key advantage, in particular if the network is large. Indeed, for a large network the number of parameters in the model increases dramatically along with the effort required to investigate the convergence assumption for every parameter if MCMC sampling is applied. These two points make INLA more attractive to use for Bayesian inference of a NMA model.

However, there are several peculiarities in many NMA models for which the implementation in `r-inla` is not straightforward. The use of correlated multivariate random effects for multiarm trials using the Kronecker product for the covariance matrix is one of the specialities which we discussed. Implementation of the node-split approach is another NMA characteristic which we could accomplish in `r-inla` using two separate likelihoods.

Node-splitting with INLA is also possible in very large networks like the application discussed by Veroniki et al. (2013) who examine inconsistency in 40 different networks with a dichotomous outcome and a total of 303 loops. INLA has a great potential for performing Bayesian inference for NMA models and offers a major alternative to MCMC software. INLA also offers possibilities for routine prior sensitivity examination (Roos and Held, 2011; Roos et al., 2015), which may be particularly useful in a NMA. We note that ranking of treatments as discussed by Lu and Ades (2006, Section 3.5) is not possible with INLA, but this approach has been recently critized by Puhan et al. (2014) as misleading since it does not take the quality of treatment effect estimates into account.

Acknowledgment We thank Martin Schumacher who suggested to investigate the possibility to perform network meta-analyses with INLA, Gerta Rücker and Milo Puhan who contributed valuable remarks and pointed to several important references and Malgorzata Roos for carefully proofreading this manuscript. Furthermore we also like to thank two anonymuous reviewer and the editor who recommended several changes which lead to substantial improvements of this paper.

Conflict of interest
The authors have declared no conflict of interest.

References

Bucher, H. C., Guyatt, G. H., Griffith, L. E. and Walter, S. D. (1997). The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *Journal of Clinical Epidemiology* **50**, 683–691.

Dias, S., Welton, N. J., Caldwell, D. M. and Ades, A. E. (2010). Checking consistency in mixed treatment comparison meta-analysis. *Statistics in Medicine* **29**, 932–944.

Fong, Y., Rue, H. and Wakefield, J. (2010). Bayesian inference for generalized linear mixed models. *Biostatistics* **11**, 397–412.

Hasselblad, V. (1998). Meta-analysis of multitreatment studies. *Medical Decision Making* **18**, 37–43.

Higgins, J. P. T. and Whitehead, A. (1996). Borrowing strength from external trials in a meta-analysis. *Statistics in Medicine* **15**, 2733–2749.

Kessels, A. G. H., Riet, G., Puhan, M. A., Kleijnen, J., Bachmann, L. M. and Minder, C. (2013). A simple regression model for network meta-analysis. *OA Epidemiology* **1**, 7.

Lu, G. and Ades, A. E. (2006). Assessing evidence inconsistency in mixed treatment comparisons. *Journal of the American Statistical Association* **101**, 447–459.

Lumley, T. (2002). Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine* **21**, 2313–2324.

Marshall, E. C. and Spiegelhalter, D. J. (2007). Identifying outliers in Bayesian hierarchical models: a simulation-based approach. *Bayesian Analysis* **2**, 409–444.

Martino, S., Akerkar, R. and Rue, H. (2011). Approximate Bayesian inference for survival models. *Scandinavian Journal of Statistics* **38**, 514–528.

Martins, T. G., Simpson, D., Lindgren, F. and Rue, H. (2013). Bayesian computing with INLA: new features. *Computational Statistics and Data Analysis* **67**, 68–83.

Mills, E. J., Ioannidis, J. P. A., Thorlund, K., Schünemann, H. J., Puhan, M. A. and Guyatt, G. H. (2012). How to use an article reporting a multiple treatment comparison meta-analysis. *Journal of the American Medical Association* **308**, 1246–1253.

Mills, E. J., Thorlund, K. and Ioannidis, J. P. A. (2013). Demystifying trial networks and network meta-analysis. *British Medical Journal* **346**, f2914.

Muff, S., Riebler, A., Held, L., Rue, H. and Saner, P. (2015). Bayesian analysis of measurement error models using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **64**, 231–252.

Norton, E. C., Miller, M. M., Wang, J. J., Coyne, K. and Kleinman, L. C. (2012). Rank reversal in indirect comparisons. *Value in Health* **15**, 1137–1140.

Paul, M., Riebler, A., Bachmann, L. M., Rue, H. and Held, L. (2010). Bayesian bivariate meta-analysis of diagnostic test studies using integrated nested Laplace approximations. *Statistics in Medicine* **29**, 1325–1339.

Plummer, M. (2003). JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. Vienna.

Puhan, M. A., Schünemann, H. J., Murad, M. H., Li, T., Brignardello-Petersen, R., Singh, J. A., Kessels, A. G. and Guyatt, G. H. (2014). A GRADE Working Group approach for rating the quality of treatment effect estimates from network meta-analysis Vienna. *British Medical Journal* **349**, g5630.

Riebler, A., Held, L. and Rue, H. (2012). Estimation and extrapolation of time trends in registry data - Borrowing strength from related populations. *The Annals of Applied Statistics* **6**, 304–333.

Roos, M. and Held, L. (2011). Sensitivity analysis in Bayesian generalized linear mixed models for binary data. *Bayesian Analysis* **6**, 259–278.

Roos, M., Martins, T. G., Held, L. and Rue, H. (2015). Sensitivity analysis for Bayesian hierarchical models. *Bayesian Analysis* **10**, 321–349.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC Press, London, UK.

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society - Series B* **71**, 319–392.

Salanti, G. (2012). Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. *Research Synthesis Methods* **3**, 80–97.

Schmidt, H. and Nehmiz, G. (2014). Joint Bayesian network meta-analysis for event counts and hazards—comparison of methods and implementations. *Value in Health* **17**, A205.

Schrödle, B., Held, L., Riebler, A. and Danuser, J. (2011). Using INLA for the evaluation of veterinary surveillance data from Switzerland: A case study. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **60**, 261–279.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* **81**, 82–86.

van Valkenhoef, G. and Ades, A. E. (2013). Evidence synthesis assumes additivity on the scale of measurement: response to "Rank reversal in indirect comparisons" by Norton et al. *Value in Health* **16**, 449–451.

van Valkenhoef, G. and Kuiper, J. (2014). *gemtc: GeMTC Network Meta-analysis*. R package version 0.6.

Veroniki, A. A., Vasiliadis, H. S., Higgins, J. P. T. and Salanti, G. (2013). Evaluation of inconsistency in networks of interventions. *International Journal of Epidemiology* **42**, 332–345.

Biometrical Journal **52** (2004) 61, zzz–zzz / DOI: 10.1002/bimj.200100000

Supplementary Material to "Network meta-analysis with integrated nested Laplace approximations"

Rafael Sauter[*,1] and **Leonhard Held** [1]

[1] University of Zurich
Epidemiology, Biostatistics and Prevention Institute
Hirschengraben 84
CH-8001 Zürich

This document contains additional material accompanying the paper "Network meta-analysis with integrated nested Laplace approximations". In Section 1 we discuss the technical details about how to implement the NMA models, presented in Section 3 and 4 in the main text, with `r-inla`. Section 2 shows the detailed results, which were used for Figure 2 in the main text, as a table. In Section 3 the parameter estimates obtained by MCMC and INLA for the application from Section 4 in the main text are shown.

1 Implementation of NMA models with the INLA R-package

The `r-inla` R-package is provided on `http://www.r-inla.org/`. The software can be installed by executing the following command lines in a R-session.

```
install.packages("INLA", repos="http://www.math.ntnu.no/inla/R/stable")
```

Information about the installed version can be retrieved by typing `inla.version()`. In order to be able to reproduce the applications from the main text make sure to install the accompanying R-package `nmainla` which is available for download as supplementary material as well (see `?install.packages`).

```
install.packages("nmainla_1.0.tar.gz")
library(nmainla)
```

The package `nmainla` essentially contains the datasets for the myocardial infarction example of Section 4.1, for the smoking cessation example of Section 4.2 and for the thrombolytic network used for the node-splitting in Section 4.3 of the main text. Detailed information for all three datasets can be obtained by typing `?myodat`, `?smokdatDI` or `?thrombdatDI`. Additionally there is a wrapper function `nodesplit_inla`, calling `r-inla` repeatedly for a list of node-splits and two additional functions `creatINLAdat` and `make.lincomb.vector` used to organise the datasets. See the `nmainla` package documentation for details about these functions and datasets.

Section 1.1 discusses the summary level data application from Section 4.1 in the main text. The implementation of a user-specified uniform prior with `r-inla`, and the settings for linear combinations covering the functional contrasts d_f are also discussed in this section. Section 1.2 discusses the implementation of the trial-arm level data discussed in Section 4.2 in the main text.

*corresponding author: e-mail: rafael.sauter@uzh.ch

The implementation of multi-arm trials with a homogeneous covariance structure is discussed in this section. Section 1.3 shows how to implement the node-splitting for the application presented in Section 4.3 in the main text.

1.1 Summary level model

The INLA results for the NMA of the myocardial infarction data discussed by Lumley (2002) was discussed in Section 4.1 in the main text.

```
require(nmainla)
data(myodat)
head(myodat)
```

```
  stre tpa atpa ret anis angio       Y  sigma     prec ind1 ind2 trtpair ind
1    1  -1    0   0    0     0 -0.0260 0.0394 644.1805    1    2       1   1
2    1   0    0   0   -1     0 -0.0048 0.0392 650.7705    1    5       2   2
3    0   1    0   0   -1     0  0.0212 0.0395 640.9229    2    5       3   3
4    1   0   -1   0    0     0 -0.1727 0.0552 328.1874    1    3       4   4
5   -1   1    0   0    0     0 -0.0684 0.0778 165.2117    2    1       1   5
6   -1   1    0   0    0     0 -0.0432 0.0634 248.7834    2    1       1   6
```

The response y_{ijk} is named `Y` in the data frame `myodat`. The observed trial variance σ^2_{ijk} is called `sigma`. `prec` is the inverse of `sigma` defining the precision σ^{-2}_{ijk}. There is a vector for the heterogeneity random effects `ind` identifying the trials (index i in the main text) and for the incoherence random effects `trtpair` (index j, k in the main text), which identifies the treatment pairs. Each trial has two treatments defined by the variables `ind1` (index j in the main text) and `ind2` (index k in the main text). The other variables (`stre, tpa, atpa, ret, anis, angio`) cover the covariates for the basic contrast parameters $\boldsymbol{d}_b$.

The `r-inla` software package requires the user to define a model formula using the typical `R` syntax (see `?formula`). The model formula will be called by the function `inla`. Additionally to the model formula there are several arguments in the `r-inla` function call which need to be specified.

The first step to implement the NMA of Section 4.1 is to define the prior distribution of the hyperparameters in the latent field. We want to use the same distributions as used by Lu and Ades (2006) and Dias et al. (2010), which is the uniform distribution $\tau \sim \mathrm{U}(0, u)$ and $\kappa \sim \mathrm{U}(0, u)$ where we set the upper bound of the distribution to $u = 10$. The implemented priors e.g. a Gamma prior on the hyperparameter can be called in `r-inla` by the argument `f(..., model = "iid", hyper(list=(prior=loggamma))`. Every latent field is initiated by `f(...)`. In this example an independent random noise latent field, named `model = "iid"` is called, which is the only latent field used for the NMA examples presented here. The `"iid"` latent model defines the prior distribution on the log-transformed precision parameter of an independent and Gaussian distributed random variable. As the uniform prior distribution is not implemented in `r-inla` we must define it. First we define the uniform distribution for the hyperparameter on the `r-inla` internal scale, which is in most of the cases the log-scale of the hyperparameter and which requests to transform the density. The transformation of densities can be done by applying the change-of-variables formula (see Held and Sabanés Bové, 2014, Appendix A.2.3)): if $f_X(x)$ is a probability density function of X one can compute the probability density function $f_Y(y)$ of the transformed variable $Y = f(X)$ by

$$f_Y(y) = f_X\{f^{-1}(y)\} \left| \frac{df^{-1}(y)}{dy} \right| = f_X(x) \left| \frac{df(x)}{dx} \right|^{-1}.$$

The latent model `r-inla` puts the prior on $\theta = f(\tau) = \log(\tau^{-2})$ and respectively on $\theta = f(\kappa) = \log(\kappa^{-2})$. We continue with the parameter τ only, but results are equivalent for κ. The uniform distribution for the interval $[0, 10]$ is $f_\tau(\tau) = 1/10$. Thus in this case we have $f^{-1}(\theta) = \sqrt{\frac{1}{\exp(\theta)}}$ and $\frac{df^{-1}(\theta)}{d\theta} = 1/2\sqrt{\exp(\theta)}$ such that

$$f_\theta(\theta) = f_\tau\{f^{-1}(\theta)\} \left| \frac{df^{-1}(\theta)}{d\theta} \right| = (1/10) \cdot 1/2\sqrt{\exp(\theta)}$$

if $\sqrt{\frac{1}{\exp(\theta)}}$ is in $[0, 10]$. This density is defined as R-function `hyperunif.function`:

```
#Upper limit for uniform distribution:
ul <- 10
#Function for Uniform distribution:
hyperunif.function <- function(x){
  if(exp(x)^-0.5 < ul & exp(x)^-0.5 > 0){
        logdens <- log(1/ul)
  }else{
        logdens <- log(0.1e-320)
  }
logdenst <- logdens+log(0.5*exp(-x/2))
return(logdenst)
}
```

The uniform distribution function `hyperunif.function` must be evaluated at a suitable grid of points, defining `prior.table`, which can then be called as prior in `r-inla`:

```
#Define grid:
lprec <-  seq(from = -40, to = 40, len = 20000)
#Create table:
prior.table <- paste(c("table:", cbind(lprec, sapply(lprec, FUN=hyperunif.function))),
                     sep = "", collapse = " ")
```

The next step is to define the NMA model-formula. We have two latent `"iid"` fields, one for heterogeneity (`ind`) and one for incoherence (`trtpair`), which defines a linear mixed models with different random effects. For both hyperparameters in each latent field we will use the uniform prior defined above as `prior.table`. The other variables for $\boldsymbol{d}_b$ enter the model as fixed effects outside `f(...)` and we exclude a global intercept, using `-1`:

```
inla.form.myo <- Y ~ -1 + tpa + atpa + ret + anis + angio +
                          f(ind, model = "iid",
                            hyper = list(theta = list(prior = prior.table))) +
                          f(trtpair , model = "iid",
                            hyper = list(theta = list(prior = prior.table)))
```

As we are eventually also interested in the functional contrasts $\boldsymbol{d}_f$ we need to define a linear combination for these, before calling the `inla(...)` function. There is a built in function `inla.make.lincombs` in the `r-inla` package which defines the linear combinations. We build a linear combination for the relative treatment difference of `ret` against `angio`.

```
LCmyo <- inla.make.lincombs(ret = 1, atpa=-1 )
names(LCmyo) <- c("ret_atpa")
```


Now the model can be estimated by calling `inla()`:

```
inla.myo <- inla(inla.form.myo, data = myodat,
                 family = "normal",
                 control.fixed = list(mean = 0, prec = 1/1000),
                 control.family = list(hyper = list(prec = list(
                                       fixed = TRUE,
                                       initial = 0))),
                 scale = prec,
                 lincomb = LCmyo,
                 control.inla = list(lincomb.derived.only = FALSE),
                 control.compute = list(dic = TRUE,cpo = TRUE)
                 )
```

The function `inla(...)` requests a model-formula like `inla.form.myo` and a dataset `data = myodat` as input. Besides the prior distribution for the hyperparameters one needs also define the fixed effect prior distribution. In `r-inla` coefficients are defined as fixed effects if they are not covered by a latent field `f(...)`. For the fixed effects `r-inla` assumes a normal distribution, which is in this case essentially the same as defining a separate latent field with `model = "iid"` but with different prior distributions.. The parameters for the normal distribution of the fixed effects can be controlled within the `inla(...)` function by the argument `control.fixed`. We again follow Lu and Ades (2006) and Dias et al. (2010) and set the prior distribution for the basic contrasts to $\boldsymbol{d}_b \sim \mathrm{N}(0, 1000)$. For normal distributions `r-inla` uses mainly precisions which is equal to the inverse of the variance. `control.fixed`. Thus we set `prec = 1/1000` and `mean = 0`.

The distribution of the response is defined by the argument `family = "normal"`. This argument is set to `binomial` for the applications of Section 4.2 and 4.3 in the main text. The `control.family` argument sets additional options which are related to the response e.g. the link function. In the summary level data example of Section 4.1 in the main text we want to scale the response variance with the observed σ^2_{ijk}. This means that we first need force the random noise which is always automatically added to the model by `r-inla` to be equal to zero. This is done by fixing the hyperparameter in `control.inla` to zero: `list(hyper = list(prec = list(fixed = TRUE, initial = 0)))`. The error can then be weighted with $1/\sigma^2_{ijk}$ by using the precision (`prec`) stored in the data frame, addressed by the argument `scale = prec`.

The linear combination which we prepared in advance are called in `inla(...)` by the argument `lincomb`. The argument `control.inla` offers many options relevant for the numerical accuracy of the approximations `r-inla` produces. The call used here `lincomb.derived.only = FALSE` increases the accuracy for the linear combination defined by `LCmyo` but generates a higher computational cost. The argument `control.compute` is useful as it offers the option to additionally compute the DIC model choice criterion. Many of the options described here have a documentation which can be accessed by typing e.g. `?control.inla`. `r-inla` generates an object of class `inla`. If one is willing to invest more computational effort one can improve the estimates for the hyperparameters by using the function `inla.hyperpar(...)` calling the `inla` object:

```
inla.myo <- inla.hyperpar(inla.myo)
```

There are several useful functions to analyse the results:

```
summary(inla.myo)
plot(inla.myo)
```

While detailed statistics about the marginal posteriors can be accessed by

```
inla.myo$summary.fixed
```


or

```
inla.myo$summary.random$trtpair
```

it is also to possible to access the marginals directly by

```
inla.myo$marginals.fixed
```

or

```
inla.myo$marginals.random$trtpair
```

which can be used as input for functions which operate on these marginals e.g. transform them. See `?inla.tmarginal`.

1.2 Trial-arm level model

The INLA results for the smoking cessation interventions network is discussed in the paper in Section 4.2. We present here the `r-inla` code for estimating a NMA model for the smoking cessation data. We especially highlight how multi-arm trials are modelled with correlated multivariate random effects.

First load the data and bring it into a suitable format by using the functions in the package `gemtc` and in the accompanying package `nmainla`.

```
#Load data (available in nmainla):
data(smokdatDI)
#Rearange data:
smokdat <- mtc.data.studyrow(data = smokdatDI,
                             armVars = c('treatment' = 't',
                                         'responders' = 'r','sampleSize' = 'n'),
                             nArmsVar = 'na' ,
                             studyNames = 1:nrow(smokdatDI),
                             patterns = c('%s', '%s%d'))
#Create baseline variable:
smokdat$baseline <- rep(smokdatDI$t1, times = smokdatDI$na)
#Study as factor:
smokdat$mu <- as.factor(smokdat$study)
#See indices suitable to INLA:
#(See function ?creatINLAdat in the nmainla package.)
smokdatINLA <- creatINLAdat(dat = smokdat,
                            treatmentvar = "treatment",
                            baselinevar = "baseline",
                            studyvar = "study")
```

The `gemtc`-function `mtc.data.studyrow` converts datasets in the one-study-per-row format to the one-arm-per-row format which is also requested by `gemtc`. The `nmainla`-function `creatINLAdat` adds indicator variables to a data frame which define the baseline contrasts ($\mathtt{d}_{1j}$), the heterogeneity random effects (`re`), for the correlated multi-arm trials the grouping vector which defines the covariance structure (`g`) and cycle-specific inconsistency random effects (`w`). The resulting data frame is suitable to use in the `inla(...)` function.

In the smoking dataset there are three basic contrasts d_{12}, d_{13} and d_{14}. Now define a model-formula for `r-inla` with correlated multivariate random effects for heterogeneity and additional trial-specific random effects for the inconsistency. This corresponds to the third model in the last plot of Figure 1 in the main text.


```
#Formula:
inla_form.smokeREinc <- responders ~ -1 + mu + d12 + d13 + d14 +
                                    f(re, model = "iid",
                                          hyper = list(theta1 = list(
                                                      prior = prior.table)),
                                                      group = g,
                                          control.group = list(model = "exchangeable",
                                          hyper = list(rho = list(fixed = TRUE,
                                                    initial = cor.inla.init)))) +
                                      f(w, model = "iid",
                                          hyper = list(theta1 = list(
                                                  prior = prior.table)))
```

The baseline contrasts are included in the model by `...+ d12 + d13 + d14 +...`. The parameter `mu` represents the baseline treatment effect a_{ij} for every trial i. The random effects for heterogeneity are included in the model similar to Section 1.1 by `f(re, model="iid", ...`. The variable `re` in the dataset `smokdatINLA` represents the trials i. We also use again the uniform prior defined in the `prior.table` object as in Section 1.1. In contrast to Section 1.1 we introduce now the grouping for the multi-arm trials by adding the argument `group = ...` inside the latent field `f(re, ...)`. We assign to the argument the vector `g` in the dataset `smokdatINLA` (`group = g`). This vector defines the position in the covariance structure $\boldsymbol{T}_i$ of every arm in the trial. The elements for the baseline treatments in `g` are set to `NA`. By the argument `control.group` we specify further how the correlation structure $\mathbf{C}_i$ looks like. We specify the model to `model = "exchangeable"` and fix the hyperparameter ρ to a certain value `cor.inla.init` by setting `fixed = TRUE`. Using `fixed = TRUE` will fix the hyperparameter at the value defined by `initial = ...` or the default value if no initial value is set. To set the initial value at $\rho = 1/2$ we must transform the ρ to the `r-inla` internal scale. In the case of the correlation parameter for multivariate random effects the internal scale for `r-inla` is Fisher's z-transformation for $\rho = 1/2$ which is

$$\log\left(\frac{1+\rho(R-1)}{(1-\rho)}\right)$$

where R is the maximal number of arms for the multi-arm trial with the most treatments in the network (Riebler et al., 2012, see section 3). So we define the transformed ρ as `cor.inla.init` as

```
# Transform group-correlation 0.5 to internal.scale of INLA:
cor <- 0.5 #correlation between treatment comparisons of the same multi-arm trial.
ngroup <- 2 #number of groups is equal to the maximum number
           #of pairwise treatment comparisons in a (multi-arm) trial.
#transformation to internal INLA-scale.
cor.inla.init <-  log((1 + cor * (ngroup - 1))/(1 - cor))
```

Besides the correlated multivariate random effects for heterogeneity we have additionally the inconsistency random effects defined by `f(w,...)` which is very similar to the definition in Section 1.1. The code example illustrates that function arguments, like for hyperparameters, inside a latent field `f(...)` must be defined as nested lists.

Now we can use again the function `inla(...)` to evaluate the model by calling the formula-object described above.

```
#Call inla:
inla.smokeREinc <- inla(as.formula(inla_form.smokeREinc),data = smokdatINLA,
                        family = "binomial", Ntrials = sampleSize,
```



```
                        control.fixed = list(expand.factor.strategy = "inla",
                                             mean = 0, prec = 1/1000),
                      )
```

As the response is binomial in this case we need to define `family = "binomial"` together with `Ntrials = ...` linking to the vector in the dataset that contains the number of events. The argument `expand.factor.strategy = "inla"` is needed to define how `r-inla` should handle the factor variable of the baseline treatments `mu`. Essentially this defines how the design matrix for this factor variable is created. See `http://www.r-inla.org/faq` for more details about this point.

The other two models shown in Figure 1 in the main text can be estimated by modifying the formula `inla_form.smokeREinc`. For the random effects model without inconsistency random effects the latent field `f(w, ...)` simply needs to be removed. For the fixed effects model both latent fields `f(re, ...)` and `f(w, ...)` can be completely removed from the formula.

1.3 Node-splitting

We present here the implementation of the node-splitting for the NMA of the thrombolytic treatment network. The INLA results for the node-splitting of the thrombolytic treatment network were compared with MCMC in Section 4.3 in the main text. To implement the node-splitting approach in `r-inla` a wrapper function `nodesplit_inla` is made available in the accompanying package `nmainla`. This function just repeatedly calls `r-inla` for a list of node-splits and organises the data such that two separate likelihoods can be applied as described in the paper.

First the data need to be loaded from the package `nmainla` and organised in a similar way as the dataset in Section 1.2 in this document.

```
#load data:
data(thrombdatDI)
# Tranform data per studyrow (gemtc-conform format):
thrombdat <- mtc.data.studyrow(data = thrombdatDI,
                               armVars = c('treatment' = 't',
                                           'responders' = 'r',
                                           'sampleSize' = 'n'),
                               nArmsVar = 'na' ,
                               studyNames = 1:nrow(thrombdatDI),
                               patterns = c('%s', '%s%d'))
# Create baseline variable:
thrombdat$baseline <- rep(thrombdatDI$t1, times = thrombdatDI$na)
# Study as factor:
thrombdat$mu <- as.factor(thrombdat$study)
#use function creatINLAdat()
#(see Rfunctions_inlaNMA.R) to add suitable indices:
thrombdatINLA <- creatINLAdat(dat = thrombdat,
                              treatmentvar = "treatment",
                              baselinevar = "baseline",
                              studyvar = "study")
```

The pairwise treatment comparison for which a node-split is possible, meaning that independent direct and indirect evidence is available in the network, was assessed by using the function `mtc.nodesplit.comparisons` available in the `gemtc` package. To do this one need to construct a network first by using `mtc.network` in the `gemtc` package which we can also plot as shown in Figure 1.


```
#network
net_thrombdat <- mtc.network(data.ab=thrombdat)
plot(net_thrombdat)
#Get node-splits:
nodecomp_thrombdat <- mtc.nodesplit.comparisons(net_thrombdat)
```

The resulting list with the treatment pairs is stored as two-column matrix in `nodecomp_thrombdat`.

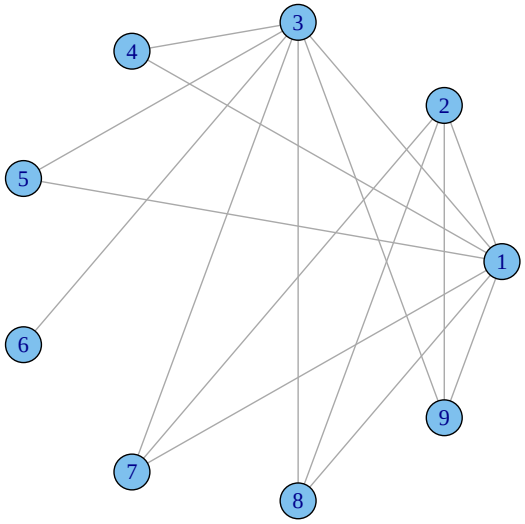

Figure 1 Plot of the network for thrombolytic trial-arm level data generated based on `net_thrombdat` which is a `mtc.network`-object defined in `R`-package `gemtc`.

Now we have all the ingredients to use `r-inla` for node-splitting with the function `nodesplit_inla`:

```
inla.node.thrombRE <- nodesplit_inla(dat = thrombdatINLA,
                                     nodelist = nodecomp_thrombdat,
                                     treatmentvar = "treatment",
                                     studyvar = "study",
                                     baselinevar = "baseline",
                                     responsevar = "responders",
                                     samplesizevar = "sampleSize",
                                     mod = "RE",
                                     priorREhyper = "hyper = list(theta1 = list(prior =
                                                  prior.table))",
                                     varf = 1000)
```

The `nmainla`-function `nodesplit_inla(...)` is a wrapper calling repeatedly `r-inla` generating the direct and indirect estimates for a given list of node-splits, defined here by `nodelist = nodecomp_thrombdat` in a NMA network. This function essentially produces the same results as the `gemtc`-function `mtc.nodesplit` which uses MCMC sampling. As described in the main text in Section 4.3, `nodesplit_inla` creates for every node-split a `inla`-formula with two separate likelihoods. A sample code for the implementation of two separate likelihoods in an other context in `r-inla` can be found on `http://www.r-inla.org/examples/case-studies/martino-akerkar-and-rue-2010` , which refers to the application discussed in Martino et al. (2011).

The option `varf` is the fixed effect prior variance for the baseline treatments and the baseline factors in the NMA-model. The option `mod` defines the NMA-model and is either equal to `FE` for a fixed effect model without random effects or equal to `RE` for a NMA-model with heterogeneity random effects. If `mod = RE` then there are three more options which should be set: `priorREhyper` defines the prior for the hyperparameters which is the variance of the heterogeneity random effects, `priorghyper` defines the prior for the grouping hyperparameters for multi-arm trials which is the correlation between multi-arm trials random effects and `cor.group` which defines the group correlation for multi-arm trials called by `priorghyper`. The `cor.group` argument must be on the `r-inla` internal scale and its default value is equal to 1.098612 which corresponds to a correlation of 1/2 which is justified by the consistency assumption. All remaining options and further descriptions can be obtained by typing `?nodesplit_inla`.

Warning: the `nodesplit_inla` function was not tested extensively and was developed for the applications presented here. Although it handles multi-arm treatments up to now it was only used for a limited number of examples and with three-arm trials only. Currently the function will not capture all possible NMA set-ups with multi-arms and thus may probably fail for some applications or produce wrong results! At the moment the function `nodesplit_inla(...)`is applicable for logistic regression NMA-models only!

2 Results for smoking cessation data models

Table 1 shows the posterior median and the 95% equi-tailed credible interval (CI) for the parameters of the three models presented in Section 4.2 in the main text. The upper part of Table 1 shows the results obtained by MCMC and the lower part the same estimates obtained by INLA. The figures of Table 1 are illustrated as forest plots in Figure 2 in the main text.

				Random effects			
		Fixed effects		heterogeneity		heterogeneity + incoherence	
		Median	95%-CI	Median	95%-CI	Median	95%-CI
MCMC:	d_{12}	0.228	-0.019 to 0.475	0.488	-0.288 to 1.306	0.448	-0.453 to 1.361
	d_{13}	0.765	0.651 to 0.879	0.835	0.391 to 1.339	0.857	0.398 to 1.394
	d_{14}	0.841	0.497 to 1.185	1.096	0.267 to 2.006	1.072	-0.012 to 2.233
	d_{23}	0.537	0.275 to 0.798	0.346	-0.466 to 1.173	0.251	-0.903 to 1.289
	d_{24}	0.612	0.239 to 0.990	0.607	-0.333 to 1.586	0.634	-0.493 to 1.791
	d_{34}	0.076	-0.262 to 0.415	0.260	-0.551 to 1.098	0.217	-0.962 to 1.318
	τ			0.818	0.550 to 1.275	0.845	0.562 to 1.311
	κ					0.567	0.033 to 4.174
	ξ_{123}					-0.080	-1.513 to 0.828
	ξ_{124}					0.006	-1.307 to 1.255
	ξ_{134}					0.001	-1.248 to 1.264
INLA:	d_{12}	0.227	-0.019 to 0.474	0.487	-0.269 to 1.274	0.439	-0.470 to 1.361
	d_{13}	0.764	0.650 to 0.879	0.833	0.397 to 1.309	0.852	0.393 to 1.375
	d_{14}	0.840	0.498 to 1.184	1.088	0.278 to 1.953	1.063	0.009 to 2.213
	d_{23}	0.537	0.276 to 0.798	0.346	-0.443 to 1.142	0.120	-2.465 to 2.153
	d_{24}	0.611	0.238 to 0.989	0.600	-0.317 to 1.545	0.623	-1.914 to 3.079
	d_{34}	0.076	-0.261 to 0.414	0.255	-0.532 to 1.059	0.232	-2.255 to 2.692
	τ			0.814	0.547 to 1.266	0.840	0.560 to 1.319
	κ					0.541	0.030 to 3.558
	ξ_{123}					-0.111	-1.550 to 0.870
	ξ_{124}					0.003	-1.430 to 1.313
	ξ_{134}					0.005	-1.320 to 1.322

Table 1 Quantiles of the marginal posterior distributions of all (baseline and functional) relative treatment effects by MCMC (top) and by INLA (bottom) for the smoking cessation data. The last lines show the estimates for the random effects variance of the incoherence τ^2 and the heterogeneity κ^2 as well as the incoherence random effect estimates ξ_{1kc}.

3 Node-splitting results for thrombolytic infarction data

One MCMC run with 100'000 iterations and a burnin of 200'000 samples was used to get the posterior distributions for all models presented in Section 4.3 in the main text. The rather large number of MCMC iterations is the same as indicated by Dias et al. (2010) motivated by the node-splitting. They mention that most of the models reach convergence for much less iterations while some of the node-split models need so many sampling iterations to satisfy the applied diagnostic convergence criteria. For MCMC sampling we rely on the `R`-package `gemtc` (van Valkenhoef and Kuiper, 2014) using `JAGS` (Plummer, 2003).

The thrombolytic treatment dataset compares 9 different treatments (1: streptokinase , 2: t-PA, 3: accelerated t-PA, 4: streptokinase and t-PA, 5: reteplase, 6: tenecteplase, 7: PTCA, 8:

urokinase, 9: anistreptilase) and reports the number of deaths in 30 or 35 days and number of patients in each treatment arm for 50 different trials. There are two three arm trials, one comparing treatments 1, 3 and 4 and one comparing treatments 1, 2 and 9. The network provides direct evidence for 16 different pairwise treatment comparisons (see also Figure 1). The pairwise treatment comparison for which a node-split is possible, meaning that independent direct and indirect evidence is available in the network, was assessed by using the function `mtc.nodesplit.comparisons` available in the `gemtc` package. As treatment 6 was only compared in one trial with treatment 3 there is no indirect evidence and thus there is no node-split possible for d_{36}. There is also no node-split for treatment 1 and 4 as there is only one other three arm trial which compares treatment 4 with treatment 3. As we assume no inconsistency within a multi-arm trial we have no other independent source of indirect evidence for d_{14}. There remain 14 possible node-splits for the direct relative treatment comparisons. The 14 possible node-splits between treatments `t1` and `t2` are contained in the data frame `nodecomp_thrombdat` produced by the `R`-code above:

```
nodecomp_thrombdat
```

```
   t1 t2
1   1  2
2   1  3
3   1  5
4   1  7
5   1  8
6   1  9
7   2  7
8   2  8
9   2  9
10  3  4
11  3  5
12  3  7
13  3  8
14  3  9
```

We use the same measure of inconsistency as proposed by Dias et al. (2010). They define the measure for the degree of inconsistency ($d_{jk}^{\text{diff.}}$) as the difference of the log-odds ratios based on direct and indirect evidence, i.e. $d_{jk}^{\text{diff.}} = d_{jk}^{\text{dir.}} - d_{jk}^{\text{ind.}}$. We thus compute the posterior distribution of $d_{jk}^{\text{diff.}}$. The result for the inconsistency estimates are shown in Table 2 for the fixed effects model and the model with random effects for heterogeneity obtained by MCMC and INLA. The results are consistent with the ones discussed in table 2 in Dias et al. (2010).

The largest inconsistency in the random and fixed effect model is found if the node for treatment 3 and 9 is split. The estimate for the marginal posterior mean of the relative treatment d_{39} based on the direct evidence is 1.36 for the random effect model by INLA. The corresponding estimate based on indirect evidence is 0.16 while the analysis based on the full data under the consistency assumption gives a marginal posterior mean estimate $\hat{d}_{39} = 0.30$ with INLA. A cross-validation approach which does not use the complete data to estimate the baseline treatment effect and heterogeneity hyperparameter yields an inconsistency estimate for the node d_{39} equal to 1.28 with a standard error equal to 3.65. The difference to the inconsistency estimate based on node splitting is with 1.21 not very large. The estimate for the standard error of the inconsistency is with node splitting only equal to 0.43 as reported in Table 2. The large difference in the uncertainty about the inconsistency between node splitting and cross validation is due to the fact that with cross-validation the heterogeneity hyperparameter τ^2 is quite different as it is estimated for the

	Fixed effects				Random effects			
	MCMC		INLA		MCMC		INLA	
	Mean	Stdev.	Mean	Stdev.	Mean	Stdev.	Mean	Stdev.
$d^{\text{diff.}}_{12}$	-0.342	0.258	-0.186	0.234	-0.342	0.258	-0.242	0.277
$d^{\text{diff.}}_{13}$	0.088	0.105	0.090	0.104	0.088	0.105	0.251	0.247
$d^{\text{diff.}}_{15}$	0.115	0.120	0.116	0.121	0.115	0.120	0.397	0.337
$d^{\text{diff.}}_{17}$	-0.273	0.219	-0.269	0.220	-0.273	0.219	-0.227	0.246
$d^{\text{diff.}}_{18}$	-0.203	0.574	-0.184	0.559	-0.203	0.574	-0.126	0.585
$d^{\text{diff.}}_{19}$	-0.453	0.255	-0.406	0.252	-0.453	0.255	-0.394	0.281
$d^{\text{diff.}}_{27}$	-0.078	0.430	-0.049	0.422	-0.078	0.430	0.017	0.450
$d^{\text{diff.}}_{28}$	-0.156	0.453	-0.143	0.446	-0.156	0.453	-0.126	0.474
$d^{\text{diff.}}_{29}$	-0.419	0.246	-0.426	0.245	-0.419	0.246	-0.510	0.290
$d^{\text{diff.}}_{34}$	-0.588	0.706	-0.649	0.668	-0.588	0.706	-0.868	0.726
$d^{\text{diff.}}_{35}$	-0.116	0.121	-0.116	0.121	-0.116	0.121	-0.397	0.337
$d^{\text{diff.}}_{37}$	0.263	0.211	0.258	0.210	0.263	0.211	0.213	0.240
$d^{\text{diff.}}_{38}$	0.284	0.463	0.267	0.453	0.284	0.463	0.216	0.484
$d^{\text{diff.}}_{39}$	1.233	0.418	1.196	0.409	1.233	0.418	1.209	0.425

Table 2 Inconsistency mean and standard deviation for all node-splits of the thrombolytic treatment network for the fixed effects and the random effects model obtained by MCMC and by INLA.

direct comparison of d_{39} only based on two observations. In the case of the fixed effect model the difference between cross-validation and node-splitting would be smaller as the heterogeneity parameter τ^2 is equal to zero. The node-split for the moderately large network of the thrombolytic treatment network with 14 node-splits and heterogeneity random effects took 192.9 minutes with 80'000 MCMC iterations which is 827 seconds per node-split or model. INLA was about 276 times faster and only used 42 seconds to complete the 14 node-splits which is 3.0 seconds per node-split. This difference in computation time is directly scalable with an increasing number of node-splits in a network.

References

Dias, S., Welton, N. J., Caldwell, D. M. and Ades, A. E. (2010). Checking consistency in mixed treatment comparison meta-analysis. *Statistics in Medicine* **29**, 932–944.

Held, L. and Sabanés Bové, D. (2014). *Applied Statistical Inference - Likelihood and Bayes*. Springer.

Lu, G. and Ades, A. E. (2006). Assessing evidence inconsistency in mixed treatment comparisons. *Journal of the American Statistical Association* **101**, 447–459.

Lumley, T. (2002). Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine* **21**, 2313–2324.

Martino, S., Akerkar, R. and Rue, H. (2011). Approximate Bayesian inference for survival models. *Scandinavian Journal of Statistics* **38**, 514–528.

Plummer, M. (2003). JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* .

Riebler, A., Held, L. and Rue, H. (2012). Estimation and extrapolation of time trends in registry data - Borrowing strength from related populations. *The Annals of Applied Statistics* **6**, 304–333.

van Valkenhoef, G. and Kuiper, J. (2014). *gemtc: GeMTC network meta-analysis*. R package version 0.6.

PAPER IV

Adaptive prior weighting in generalized linear models

Leonhard Held, Rafael Sauter

Paper in revision for *Biometrics*.

Adaptive prior weighting in generalized regression

Leonhard Held and Rafael Sauter

Department of Biostatistics

Epidemiology, Biostatistics and Prevention Institute

University of Zurich

Hirschengraben 84, 8001 Zurich, Switzerland

Email: leonhard.held@uzh.ch, rafael.sauter@uzh.ch

SUMMARY: The prior distribution is a key ingredient in Bayesian inference. Prior information on regression coefficients may come from different sources and may or may not be in conflict with the observed data. Various methods have been proposed to quantify a potential prior-data conflict, such as Box's p-value. However, are no clear recommendations how to react to possible prior-data conflict in generalized regression models. To address this deficiency, we propose to adaptively weight a pre-specified multivariate normal prior distribution on the regression coefficients. To this end, we relate empirical Bayes estimates of prior weight to Box's p-value and propose alternative fully Bayesian approaches. Prior weighting can be done for the joint prior distribution of the regression coefficients or - under prior independence - separately for pre-specified blocks of regression coefficients. We outline how the proposed methodology can be implemented using integrated nested Laplace approximations (INLA) and illustrate the applicability with a Bayesian logistic regression model for data from a cross-sectional study. We also provide a simulation study that shows improved performance of our approach in the case of prior misspecification in terms of root mean squared error and coverage. Supplementary material gives details on software code and another application to a Bayesian analysis of binary longitudinal data from a randomized clinical trial using a generalized linear mixed model.

.

KEY WORDS: Hyper-g prior; Prior weight; prior-data conflict; g-prior; generalized regression; INLA

1. Introduction

Appropriate specification of the prior distribution is a key ingredient in Bayesian statistics. It is also considered as the most controversial feature of Bayesian inference. In this paper we discuss the role of the prior distribution in regression models from a novel perspective. We consider a commonly used setup where a proper multivariate normal prior is assigned to the regression coefficients. Prior weighting is achieved by a scalar $g > 0$, acting multiplicatively on the prior covariance matrix. Thus, the prior weight is represented by the inverse $w = 1/g$. The focus of this paper will be on empirical and fully Bayesian approaches to estimate the inverse prior weight g from the data at hand.

We distinguish four different sources for a prior distribution. First, prior information may come from historical data of the same structure as the current data. For example, data from past clinical trials may be used to construct a suitable prior for the analysis of data from a current trial with the same outcome. Approaches to integrate historical data include the robust meta-analytic approach (Schmidli et al., 2014) and the power prior (Ibrahim and Chen, 2000; Duan et al., 2006; Neuenschwander et al., 2009), which introduces a weight parameter to discount historical data.

Secondly, the prior distribution may come from elicitation of expert opinion (O'Hagan et al., 2006; Spiegelhalter et al., 2004, Section 5.2). For example, Miettinen et al. (2008) develop a risk prediction model for the presence of pneumonia, elicited from 22 clinical experts. This model has been subsequently updated in Held et al. (2012) using data on more than 600 patients presenting with cough and fever at a general practitioner's practice in Switzerland.

However, historical data or expert opinion may not be available for the problem at hand, but an informative prior may still be warranted based on contextual reasoning. Greenland (2006, 2007a,b, 2009) argues strongly that proper priors should be used in

the analysis of epidemiological studies to avoid implicit unrealistic assumptions of the corresponding frequentist analysis (operationally equivalent to a Bayesian analysis with improper priors on the parameters of interest). For example, Greenland (2006) specifies a normal prior with mean zero and variance 1/2 for a log odds ratio parameter to reflect the prior belief that the median odds ratio is 1 and the odds ratio is between 1/4 and 4 with 95% probability *a priori*. Other choices for prior mean and variance are possible, of course, and Greenland (2006) recommends to perform a sensitivity analysis by varying the prior variance.

Fourthly and finally, proper default prior distributions may be used as a conservative guess or to avoid the problem of diverging maximum likelihood estimates in logistic regression due to complete separation (Albert and Anderson, 1984). For example, the ridge prior, with prior mean zero and prior covariance matrix proportional to the identity matrix, is a commonly used default prior. Zellner's g-prior for linear models (Zellner, 1986), with prior covariance matrix proportional to the covariance matrix of the maximum likelihood estimate (MLE) of the regression coefficients, is another default prior, which has the attractive feature that g can be interpreted as relative inverse prior sample size, see for example Marin and Robert (2007, Section 3.2.2) or Liang et al. (2008). The g-prior is a natural approach to incorporate prior correlations between regression coefficients (see the application described in supplementary material) and automatically adjusts for different variances of the covariates. Suitable extensions of the g-prior to generalized linear models (GLMs) are discussed in Sabanés Bové and Held (2011). Both ridge and g-priors are often used for Bayesian model selection, where the prior distribution needs to be proper to ensure that the marginal likelihood is well-defined and the corresponding Bayes factors (Kass and Raftery, 1995) can be calculated.

Methodology to estimate the inverse prior weight g goes back to the literature on ridge

regression (Lindley and Smith, 1972; Hoerl et al., 1975; Box, 1980). Empirical Bayes (EB) estimates of g in the context of g-priors have been proposed by Copas (1983) both for the linear and the logistic regression model. Fully Bayesian (FB) approaches to estimate g have been advocated in the linear model with regression splines (Denison et al., 2002, Section 3.5 and references in Section 3.8), using an inverse gamma hyperprior for g in combination with a ridge prior. The support of the inverse gamma distribution is the whole positive real line, thus the prior weight can be either de- or increased. Prior distributions for the parameter g of the g-prior have been proposed in Cui and George (2008); Liang et al. (2008) and Held et al. (2015) in the context of Bayesian model selection.

This paper is structured as follows. In the generalized linear model with a multivariate normal prior on the regression coefficients (Section 2) we first discuss methodology originally proposed by Box (1980) to quantify the prior-data conflict, see also Spiegelhalter et al. (2004, Section 5.8), Greenland (2006) and Evans and Moshonov (2006). We then proceed and describe methods to estimate the prior weight, represented by the parameter $1/g$. This leads to *adaptive* prior weighting, as opposed to approaches with fixed prior weight. We review empirical Bayes procedures (Copas, 1983, 1997) to estimate g in the g-prior setting and extend those to any normal prior. Furthermore, we show that EB estimates of g correspond to intermediate solutions between prior-data agreement and disagreement. We finally propose fully Bayes procedures to estimate g using a suitable hyperprior for g. If blocks of regression coefficients are *a priori* independent, then the approach can be extended to separately weight each block. Application in more complex regression models is also possible, for example in generalized linear mixed models. Inference for Bayesian GLMs with a hyperprior on g is done using integrated nested Laplace approximations (INLA) (Rue et al., 2009), to avoid the commonly used computer-intensive Markov chain Monte Carlo (MCMC) simulation.

Two applications are considered in this paper: A Bayesian logistic regression model for data from a cross-sectional study (Sullivan and Greenland, 2013) is described in Section 3.1, while a Bayesian analysis of binary longitudinal data with a generalized linear mixed model is outlined in supplementary material. Section 3.2 describes additional simulation studies that have been performed to investigate the properties of the proposed methodology.

2. Methodology

Consider a generalized linear model (GLM) with outcomes y_i, $i = 1, \ldots, n$, and linear predictor $\eta_i = \alpha + \boldsymbol{x}_i^\top \boldsymbol{\beta}$, where the vector of regression coefficients $\boldsymbol{\beta}$ has dimension d. The mean $\mu_i = h(\eta_i)$ of y_i is obtained with the response function $h(\eta_i)$, the variance function $v(\mu_i)$ determines the variance of y_i. We use a Gaussian prior with mean $\boldsymbol{\nu}$ and covariance matrix $g\,\boldsymbol{\Sigma}$ for $\boldsymbol{\beta}$, *i.e.* $\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{\nu}, g\,\boldsymbol{\Sigma})$. The intercept α can be extremely sensitive to how covariates are centered and how factors are coded, so we follow the recommendations by Greenland and Mansournia (2015, Section 7) and use Jeffreys' prior $f(\alpha) \propto 1$. More informative priors may induce unjustifiable shrinkage of the intercept towards an arbitrary prior mean. We note that also Gelman et al. (2008) use an extremely dispersed Cauchy prior for the intercept, negligibly different from our flat prior.

2.1 *Prior-data conflict*

Box (1980) has suggested an approach to quantify a potential conflict between the prior distribution and the observed data. The methodology is based on the prior predictive distribution $f(\boldsymbol{y})$ of the data $\boldsymbol{Y}$ and compares the distribution of $f(\boldsymbol{Y})$ with $f(\boldsymbol{y})$, evaluated at the observed data $\boldsymbol{y} = \boldsymbol{y}_{\text{obs}}$. Box's p-value is based on the probability

$$\Pr\{f(\boldsymbol{Y}) \leqslant f(\boldsymbol{y}_{\text{obs}})\}, \tag{1}$$

where a small value of (1) implies that the observation $\boldsymbol{y}_{\text{obs}}$ has relatively low prior predictive density, *i. e.* indicates prior-data conflict. To avoid some anomalous behavior, Evans and Moshonov (2006) proposed to replace $\boldsymbol{Y}$ in (1) with a minimal sufficient statistic for the parameter of interest. This ensures that the method provides a measure of prior-data conflict only, and not a confounded check of the model + prior combination. In more recent work, Evans and Jang (2011a) show the consistency of the Evans and Moshonov (2006) methodology and discuss the lack of invariance of the original Box (1980) approach, see also Evans and Jang (2010, 2011b).

However, exact computation of (1) is difficult in GLMs. Therefore, Greenland (2006) suggested to consider the MLE $\hat{\boldsymbol{\beta}}_{\text{ML}}$ (of course a minimal sufficient statistic for $\boldsymbol{\beta}$) as the "data" with (asymptotic) $\hat{\boldsymbol{\beta}}_{\text{ML}} \,|\, \boldsymbol{\beta} \sim \text{N}(\boldsymbol{\beta}, \boldsymbol{\mathcal{T}})$ distribution (Fahrmeir and Kaufmann, 1985), where $\boldsymbol{\mathcal{T}}$ denotes the (estimated) covariance matrix of the MLE. Combining this with a $\text{N}(\boldsymbol{\nu}, g\,\boldsymbol{\Sigma})$ prior for $\boldsymbol{\beta}$ gives the (approximate) prior predictive distribution $\hat{\boldsymbol{\beta}}_{\text{ML}} \sim \text{N}(\boldsymbol{\nu}, \boldsymbol{\mathcal{T}} + g\,\boldsymbol{\Sigma})$. The standardized difference

$$T(g) = (\hat{\boldsymbol{\beta}}_{\text{ML}} - \boldsymbol{\nu})^{\top} (\boldsymbol{\mathcal{T}} + g\,\boldsymbol{\Sigma})^{-1} (\hat{\boldsymbol{\beta}}_{\text{ML}} - \boldsymbol{\nu}) \tag{2}$$

can then be evaluated against a χ^2-distribution with d degrees of freedom to compute Box's p-value. This approximates the predictive check by Box (1980, eq. (3.9)) for the linear model based on the F-distribution using an additional improper prior $f(\sigma^2) \propto \sigma^{-2}$ on the residual variance.

2.2 *Estimates of prior weight*

In the absence of prior information on $\boldsymbol{\Sigma}$, the generalized g-prior (Sabanés Bové and Held, 2011) can be used as default. The corresponding prior covariance matrix is taken as $\boldsymbol{\Sigma} = c\,(\boldsymbol{X}^{\top}\boldsymbol{W}\boldsymbol{X})^{-1}$ where $\boldsymbol{W}$ is a diagonal matrix with corresponding weights on the diagonal (*e. g.* the binomial sample sizes for logistic regression). Here, the columns of

the design matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^\top$ are assumed to be centred, *i. e.* $\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{1} = \boldsymbol{0}$. The constant $c = c(\alpha)$ depends on the specific GLM and is defined as

$$c(\alpha) = v(h(\alpha))\{h'(\alpha)\}^{-2} \tag{3}$$

where $h'(.)$ is the derivative of $h(.)$, see Copas (1983) for a derivation for binary outcomes and Sabanés Bové and Held (2011) for a general treatment. Under the generalized g-prior, the implied shrinkage of the MLE $\hat{\boldsymbol{\beta}}_{\text{ML}}$ is approximately as in the linear model (Held et al., 2015) with posterior mean

$$\mathsf{E}(\boldsymbol{\beta} \,|\, \boldsymbol{y}) \approx \left(\frac{n \cdot \hat{\boldsymbol{\beta}}_{\text{ML}} + n/g \cdot \boldsymbol{\nu}}{n + n/g} \right),$$

which reduces to

$$\mathsf{E}(\boldsymbol{\beta} \,|\, \boldsymbol{y}) \approx \frac{g}{g+1} \hat{\boldsymbol{\beta}}_{\text{ML}} \tag{4}$$

for $\boldsymbol{\nu} = \boldsymbol{0}$. Thus $t = g/(g+1)$ can be interpreted as shrinkage factor for the generalized g-prior with prior mean $\boldsymbol{0}$.

To derive an EB estimate of g, we note that we can re-write (2) with $\boldsymbol{\mathcal{T}} \approx c\,(\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1}$ (Copas, 1983) as

$$T(g) = \frac{1}{1+g} \frac{1}{c} (\hat{\boldsymbol{\beta}}_{\text{ML}} - \boldsymbol{\nu})^\top (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}) (\hat{\boldsymbol{\beta}}_{\text{ML}} - \boldsymbol{\nu}). \tag{5}$$

Equating (5) with its expectation d (subject to $g \geqslant 0$) gives the analytic solution

$$\begin{aligned} \hat{g} &= \max \left\{ \frac{1}{d} \frac{1}{c} (\hat{\boldsymbol{\beta}}_{\text{ML}} - \boldsymbol{\nu})^\top (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}) (\hat{\boldsymbol{\beta}}_{\text{ML}} - \boldsymbol{\nu}) - 1, 0 \right\} \\ &\approx \max\{z_{\text{obs}}/d - 1, 0\}, \end{aligned} \tag{6}$$

here z_{obs} denotes the observed deviance (relative to the null model with $\boldsymbol{\beta} = \boldsymbol{0}$), obtained from fitting a standard GLM (Copas, 1983, 1997) to the data at hand. By construction, plugging-in (6) into (5) gives (for $\hat{g} > 0$) $T(\hat{g}) = d$, so Box's P-value can be easily evaluated for the adapted $\mathrm{N}(\boldsymbol{\nu}, \hat{g}\boldsymbol{\Sigma})$ prior with EB estimate $\hat{g}$. Box's P-value turns out

to be $0.32, 0.37, 0.39 \to 0.5$ for increasing degrees of freedom $d = 1, 2, 3 \to \infty$. This illustrates that in regular cases (where $\hat{g} > 0$) the empirical Bayes approach to estimate g is a way to avoid extreme prior-data agreement and disagreement with unremarkable Box's P-values between 0.32 and 0.5. If $\hat{g} = 0$ then Box's P-value will be even larger.

The approach can be easily extended to arbitrary prior mean $\boldsymbol{\nu}$ if we evaluate the deviance not against the null model $\boldsymbol{\nu} = \mathbf{0}$ but against a model with non-zero prior mean $\boldsymbol{\nu}$. This can be achieved by fitting a GLM with offset $\boldsymbol{X\nu}$. For arbitrary prior covariance matrix $\boldsymbol{\Sigma}$ an empirical Bayes-type (moment-based) estimate of g can be implemented by equating (2) with the mean d of the $\chi^2(d)$-distribution and numerically solving for g.

The empirical Bayes approach avoids arbitrary choices of g which may be at odds with the data. However, the uncertainty about the estimate $\hat{g}$ is ignored, *i. e.* the estimate $\hat{g}$ is treated as the true value g. This is particularly worrying if $\hat{g} = 0$, since then the posterior of $\boldsymbol{\beta}$ degenerates to a point mass at the prior mean $\boldsymbol{\nu}$, no matter what the data are. In contrast, a fully Bayesian approach to estimate g will incorporate the uncertainty about the estimate from its posterior distribution. If the prior distribution comes from historical data, a beta prior is commonly used for $1/g$, which restricts the range of g to values larger than unity (Duan et al., 2006). The prior can therefore only be down- but not up-weighted. However, if the prior distribution is not based on historical data, then also increasing the weight of the prior distribution may be warranted by the data at hand. We will illustrate this in Application 3.1.

For Bayesian model selection based on the g-prior, Liang et al. (2008) suggest to use the hyper-g prior with prior density

$$f(g) = \frac{a-2}{2}(1+g)^{-a/2} \tag{7}$$

for g, which is proper for $a > 2$. This prior distribution is a special case of a class of prior distributions proposed by Cui and George (2008) and induces a beta distribution for the

shrinkage factor $t = g/(g+1)$: $t \sim \text{Be}(1, a/2-1)$. Of particular interest is the case $a = 4$, where the prior on the shrinkage factor t is standard uniform and thus the prior median of g is 1. Furthermore, the distribution of $w = 1/g$ is the same as the distribution of g, *i. e.* the prior has no preference regarding up- or down-weighting. The cdf of g has a simple analytic form, $F(g) = g/(g+1)$, so prior probabilities of interest can be easily calculated, *e. g.* $\Pr(1/2 \leqslant g \leqslant 2) = 1/3$ or $\Pr(1/19 \leqslant g \leqslant 19) = 0.9$. We consider this "standard" hyper-g prior as sufficiently dispersed since g has infinite expectation. Furthermore, under the generalized g-prior a uniform prior on the shrinkage factor t implies that the posterior mode of the shrinkage factor $t = g/(g+1)$ is asymptotically equal to the corresponding EB estimate based on (6) (Held et al., 2015). Thus, the standard hyper-g prior regularizes empirical Bayes and can be considered as a natural choice for a hyperprior for g.

An alternative symmetric prior would be $f(g) = \pi^{-1} g^{-0.5} (1+g)^{-1}$, which corresponds to $t \sim \text{Be}(1/2, 1/2)$, the so-called horseshoe prior (Carvalho et al., 2010). This choice is also indifferent regarding up- or down-weighting, but puts substantially more prior mass to extreme values of g. For example, under the horseshoe prior $\Pr(1/161 \leqslant g \leqslant 161) \approx 0.9$.

The Strawderman-Berger (short Strawderman) prior (Berger, 1980), obtained from (7) with $a = 3$, places more weight on larger values of g, *i. e.* treats g not symmetric. For example, the prior median is 3. Another non-symmetric prior on g is inverse gamma distribution $\text{IG}(a, b)$ with mode $b/(a+1)$. This choice is often made for convenience due to conjugacy in the normal linear model (Denison et al., 2002, Section 3.3), but lacks a deeper motivation as a suitable prior for the weight parameter $w = 1/g$.

Whatever prior for g is used, calculation of the posterior distribution can be done using

numerical integration with INLA (Rue et al., 2009), which we describe in Section 2.3. We also comment briefly on alternative MCMC procedures.

2.3 *Implementation in INLA*

The traditional choice to implement the proposed approach would be Markov chain Monte Carlo (MCMC), but we prefer a numerical approach based on INLA to avoid potential convergence problems and the associated Monte Carlo error of MCMC. However, adaptive prior weighting with the `R-INLA` interface (see Martins et al. (2013, Section 2.3) for a summary) requires specific amendments to the model which we now briefly describe. More details can be found in Supplementary Material.

Consider a GLM as described in Section 2 with linear predictor $\eta_i = \alpha + \boldsymbol{x}_i^\top \boldsymbol{\beta}$ where $\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{\nu}, g\boldsymbol{\Sigma})$ *a priori*. `R-INLA` does not allow to specify this prior directly, so the formulation needs to be re-written as $\eta_i = \alpha + \boldsymbol{x}_i^\top \boldsymbol{\nu} + \boldsymbol{x}_i^\top \widetilde{\boldsymbol{\beta}}$ where $\widetilde{\boldsymbol{\beta}} \sim \mathrm{N}(\mathbf{0}, g\boldsymbol{\Sigma})$. Therefore, $o_i = \boldsymbol{x}_i^\top \boldsymbol{\nu}$ can be used as offset variable and it is sufficient to consider priors for the regression coefficients $\widetilde{\boldsymbol{\beta}}$ with mean zero. The idea is to treat the mean-zero regression coefficients $\widetilde{\boldsymbol{\beta}}$ as a Gaussian Markov random field (GMRF) (Rue and Held, 2005) with pre-specified precision matrix $\boldsymbol{\Sigma}^{-1}$ - up to the possibly unknown multiplicative weight factor $w = 1/g$. However, it is not possible to directly compute the product $\boldsymbol{x}_i^\top \widetilde{\boldsymbol{\beta}}$ for all observations $i = 1, \ldots, n$. The trick is now to use the `copy` feature (Martins et al., 2013, Section 4.3) in order to define d identical copies of $\widetilde{\boldsymbol{\beta}}$ in the model formulation, eventually multiplied with the covariate values $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{nj})^\top$, $j = 1, \ldots, d$.

The `R-INLA` default treats the weight $w = 1/g$ as unknown and uses a gamma hyperprior for it. However, the software allows the user to define any suitable prior density, either as an expression or in tabulated form, given the value of the prior density on a suitable grid. Note that INLA requires this to be done for $\log(w)$. Here we have used the tabulated approach for the standard hyper-g, horseshoe and Strawderman prior and

have computed the corresponding density of $\log(1/g)$ with a change-of-variables. The weight w can also be fixed at any pre-specified value which we have used to treat the case $g = 1$.

The described implementation can be generalized to independent GMRFs $\widetilde{\boldsymbol{\beta}}_1, \ldots, \widetilde{\boldsymbol{\beta}}_p$, say, in order to weight the corresponding components of $\boldsymbol{\beta}$ with separate weight parameters $g_1, \ldots, g_p$, see end of Section 3.1 for an example. It is also straightforward to adaptively weight the prior on the regression coefficients $\boldsymbol{\beta}$ in more complex models such as generalized linear mixed models, see the application described in Supplementary Material.

A possible implementation with MCMC would combine the Gamerman (1997) algorithm for Bayesian generalized linear models (for fixed g) with a Metropolis-Hastings sampler from the full conditional of g. This can be easily and efficiently implemented, but the analysis with INLA is still much faster and provides estimates without Monte Carlo error. This is particularly important for the simulation studies reported in Section 3.2, where we fit hundred thousands of different prior-data combinations.

3. Applications

3.1 *Bayesian analysis of a logistic regression model*

Sullivan and Greenland (2013) consider data from a cross-sectional study on obstetric care and neonatal death at a teaching hospital. The binary outcome variable (death yes/no) is related to 14 explanatory variables. They are all binary with frequencies between 0.3% (variable `hydram`) and 77% (variable `nomonit`). There are only 17 deaths observed among 2992 births. Sullivan and Greenland (2013) give more information about the data originally from Neutra et al. (1978) and select an informative prior for $\boldsymbol{\beta}$ for a Bayesian logistic regression analysis. The corresponding regression coefficient vector $\boldsymbol{\beta}$

is assumed to be *a priori* normally distributed with mean $\boldsymbol{\nu}_{\text{SG}} = \log(\mathbf{OR})$, where the vector $\mathbf{OR} = (2,2,2,4,2,1,4,2,2,2,4,2,2,4)^\top$ contains the prior median odds ratios for each explanatory variable. The prior covariance matrix $\boldsymbol{\Sigma}$ has been chosen to be diagonal with all variances equal to $1/2$. The intercept is considered as a nuisance parameter with improper (flat) prior.

There is no evidence for prior-data disagreement for the informative $\text{N}(\boldsymbol{\nu}_{\text{SG}}, \boldsymbol{\Sigma})$ prior (equation (2) gives Box's p=0.91). Even if the prior mean $\boldsymbol{\nu}$ would be set to zero, *i.e.* for a $\boldsymbol{\beta} \sim \text{N}(\mathbf{0}, \boldsymbol{\Sigma})$ prior, there would be no compelling evidence for prior-data conflict (Box's p=0.13). The EB estimates of g are $\hat{g} = 0.00$ and $\hat{g} = 2.10$ in these two cases. Box's p-values using the EB estimates of g in the prior covariance matrix $\hat{g}\boldsymbol{\Sigma}$ are p=0.60 and p=0.45, respectively, so in both cases close to 0.5, as expected from the discussion in Section 2.2.

If we combine the Sullivan and Greenland (2013) prior with the standard hyper-g prior (7) (with $a = 4$), the resulting posterior for g (see Figure 1) has median 0.16 (equi-tailed 95% credible interval (CI): 0.01 to 0.81). Thus, the hyper-g prior increases the weight of the prior on the regression coefficients by a median factor of $1/0.16 \approx 6.3$, which corresponds to a reduction of the prior variance from 0.5 to 0.08. However, there is quite large uncertainty regarding g where values larger than 1 still have some posterior mass.

Sullivan and Greenland (2013) pay particular attention to one explanatory variable, `hydram` (x_7), an indicator of hydramnios during pregnancy, with MLE 60 (95% Wald confidence interval 5.7 to 635, profile likelihood confidence interval 2.8 to 478) of the corresponding odds ratio $\exp(\beta_7)$. The Sullivan and Greenland (2013) prior for the corresponding log odds ratio β_7 is normal with mean $\log(4)$ and variance $1/2$, resulting in a prior median odds ratio of 4 with equi-tailed 95% prior CI from 1 to 16. Combining this prior with the data gives posterior median 6.1 (95% CI: 1.6 to 22.8) for the odds ratio.

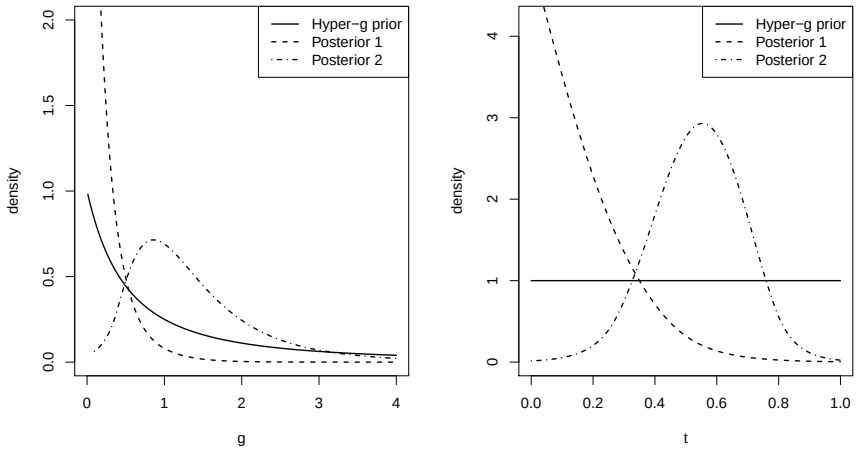

Figure 1: Hyper-g prior and posterior density of g (left) and the corresponding $t = g/(1+g)$ (right) in the logistic regression example. Posterior 1 based on $\mathrm{N}(\boldsymbol{\nu}_{\mathrm{SG}}, g\,\boldsymbol{\Sigma})$ prior for $\boldsymbol{\beta}$. Posterior 2 based on $\mathrm{N}(\mathbf{0}, g\,\boldsymbol{\Sigma})$ prior for $\boldsymbol{\beta}$.

If we treat g as unknown with hyper-g prior, then the posterior median of $\exp(\beta_7)$ is 4.3 (95% CI: 2.3 to 10.5). Although the hyper-g prior implies a substantially more dispersed marginal prior on the odds ratio (95% prior CI: 0.21 to 75.0), the posterior is actually narrower than for fixed $g = 1$. The corresponding OR estimates under the hyper-g, horseshoe and Strawderman prior are given in Table 1, together with DIC values to assess the model fit. The posterior distributions of the corresponding regression coefficient β_7 (*i.e.* the log odds ratio) are compared in Figure 2. Of note, the posterior under the horseshoe prior is substantially more peaked and narrower than under the hyper-g and Strawderman prior. The model fit turns out to be 3-4 units better for the three fully Bayesian approaches compared to the analysis with fixed $g = 1$, with the horseshoe prior having the lowest DIC value.

The same analysis has been conducted with prior mean $\boldsymbol{\nu} = \mathbf{0}$ and the same diagonal prior covariance matrix $\boldsymbol{\Sigma}$. The resulting posterior of g has median 1.20 (95%-CI: 0.46 to 3.13) and is also displayed in Figure 1. Thus, the hyper-g prior now decreases the

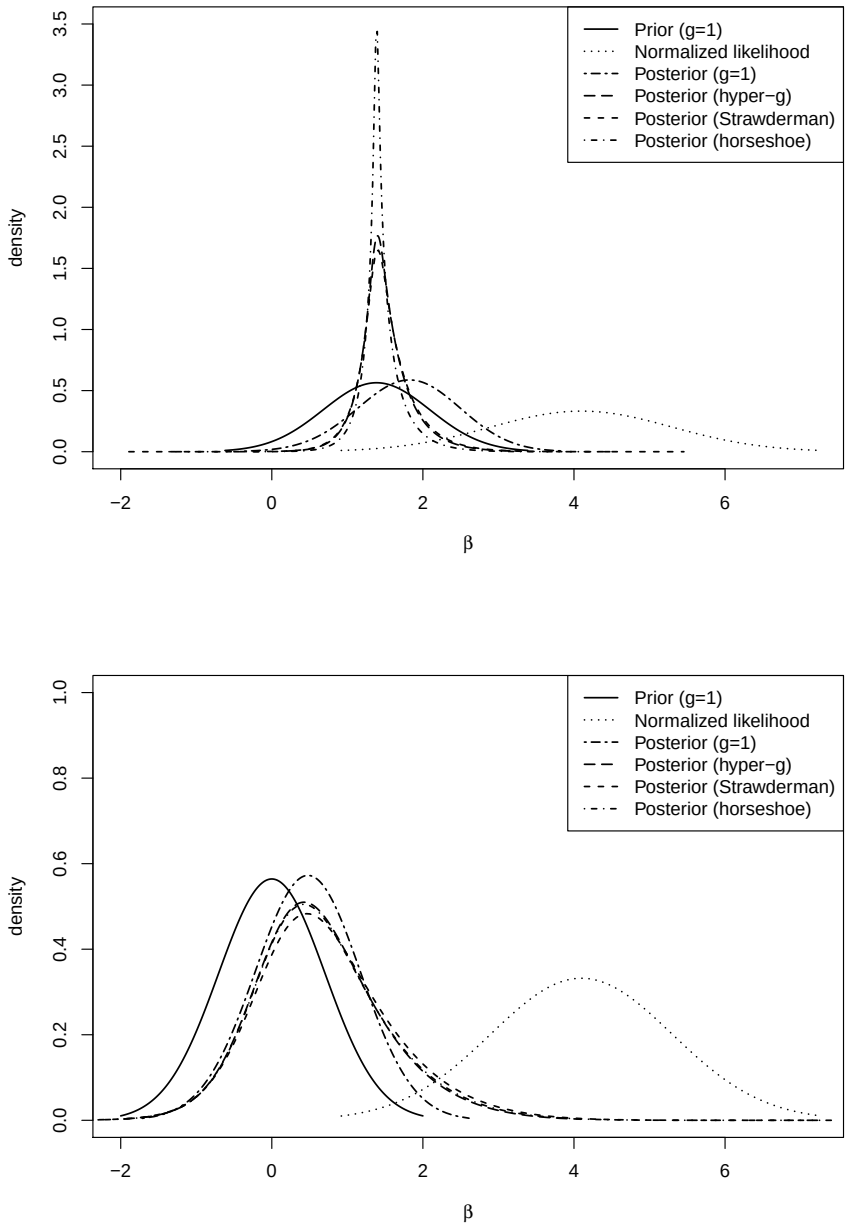

Figure 2: Posterior density of regression coefficient β_7 for the variable `hydram` in the logistic regression example with fixed ($g = 1$) and adaptive (hyper-g) prior weighting. Top: $\mathrm{N}(\boldsymbol{\nu}_{\mathrm{SG}}, g\,\boldsymbol{\Sigma})$ prior. Bottom: $\mathrm{N}(\mathbf{0}, g\,\boldsymbol{\Sigma})$ prior.

weight of the prior on the regression coefficients, but only slightly by a median factor of $1/1.20 = 0.83$, with considerable uncertainty regarding g. Accordingly, the posterior distribution of the regression coefficient related to the variable `hydram` now barely differs whether we use fixed $g = 1$ or the hyper-g prior, see Figure 2. Indeed, the posterior median of $\exp(\beta_7)$ is now 1.6 (95% CI: 0.4 to 6.3) for $g = 1$ and 1.8 (95% CI: 0.4 to 13.4) for the hyper-g prior, estimates for the other two priors (horseshoe and Strawderman-Berger) are given in Table 1, again with DIC values, which are now very similar for the different approaches.

Figure 1 also shows prior and posterior of $t = g/(g+1)$. For the prior mean $\boldsymbol{\nu}_{\text{SG}}$, the posterior mode of t is close to zero, as expected from the EB estimate $\hat{g} = 0$. For prior mean $\mathbf{0}$, the posterior mode of t is 0.55, slightly smaller than the corresponding EB estimate $\hat{g} = 2.10/(1+2.10) = 0.68$. The difference can be explained by the substantial non-normality of the posterior distribution of $\boldsymbol{\beta}$, *cf*. the difference between the Wald and the profile likelihood confidence interval for β_7 given above. Also, the correspondence of the EB and FB estimates has been shown only for the (generalized) g-prior, which is not used here.

One could argue that the above change to the prior mean $\boldsymbol{\nu} = \mathbf{0}$ should be accompanied by a more flexible formulation for the prior variances. To do so, we now introduce a new prior "block hyper-g" formulation with three different g parameters: g_1 for the block of nine covariates with original prior mean of $\log(2)$, g_2 for the block of four covariates with prior mean of $\log(4)$, and g_3 parameter for the single covariate with prior mean 0. Thus the prior weight is now allowed to vary from block to block.

The posterior median of g_1 is 0.19 (95% CI: 0.02 to 1.4), g_2 has posterior median 4.1 (95% CI: 0.9 to 20.9), while g_3, the inverse prior weight of the single covariate with prior mean 0, has posterior median 0.64 with large posterior uncertainty (95% CI: 0.04

to 9.6). Thus, the weight of the prior distribution has been increased by a median factor of $1/0.19 \approx 5.3$ for the first block of parameters, whereas the weight of the second block (which includes the variable `hydram`) has been decreased by a median factor of 4.1. For example, the posterior median of $\exp(\beta_7)$ is now 5.5 (95% CI: 0.4 to 95.5). Thus, the decreased weight of the prior distribution leads to a substantially larger OR estimate and a decreased precision of the regression coefficient, compared to the analysis with one unknown weight parameter $1/g$. Of note, this formulation gives the best model fit with DIC value 182.6, see Table 1.

	$\boldsymbol{\nu} = \boldsymbol{\nu}_{\mathbf{SG}}$			$\boldsymbol{\nu} = \mathbf{0}$		
	OR	95% CI	DIC	OR	95% CI	DIC
ML	60	5.7 to 634.7		60	5.7 to 634.7	
g=1	6.1	1.6 to 22.8	183.4	1.6	0.4 to 6.3	188.3
Strawderman	4.3	2.3 to 11.1	180.5	1.9	0.4 to 16.1	188.8
Hyper-g	4.3	2.3 to 10.5	180.3	1.8	0.4 to 13.4	188.8
Horseshoe	4.1	2.6 to 8.4	179.6	1.8	0.4 to 14.4	188.9
block Hyper-g				5.5	0.4 to 95.5	182.6
Prior (g=1)	4.0	1.0 to 16.0		1.0	0.25 to 4.00	

Table 1: Odds ratio (OR) estimate and 95% credible interval for hydramnios coefficient with prior mean $\boldsymbol{\nu}_{\mathrm{SG}}$ (left) or prior mean $\mathbf{0}$ (right) together with DIC for different priors on g.

3.2 *Simulation studies*

In a simulation study we have compared our approach with different hyperpriors for g (including fixed $g = 1$) and different degrees of misspecification of the prior mean (Section 3.3) or the covariance matrix (Section 3.4). To do so, we simulate $\boldsymbol{\beta}$ from a (possibly misspecified) "prior" distribution and subsequently $\boldsymbol{y}$ from a logistic regression model with linear predictor $\alpha + \mathbf{X}^\top \boldsymbol{\beta}$, here $\boldsymbol{X}$ is the same design matrix as in the application described in Section 3.1. For the subsequent analyses with INLA we use a normal prior for $\boldsymbol{\beta}$ with mean $\boldsymbol{\nu}_{\mathrm{SG}}$ and covariance matrix $g\boldsymbol{\Sigma}$ where $\boldsymbol{\Sigma} = \mathrm{diag}(0.5, \ldots, 0.5)$.

3.3 *Simulation study I with shifted mean*

Misspecification of the prior mean $\boldsymbol{\nu}_{\text{SG}}$ is achieved by adding a shift parameter ϵ_s, $s = 1, \ldots, S$ to each component of $\boldsymbol{\nu}_{\text{SG}}$. Here we use $\epsilon_s \in \boldsymbol{\epsilon} = (-2.6, -2.4, \ldots, 0, \ldots, 2.4, 2.6)$, so $S = 27$, and sample $\boldsymbol{\beta}_s^{(k)}$, $k = 1, \ldots, K = 1000$ from $\text{N}(\boldsymbol{\nu}_{\text{SG}} + \epsilon_s, \boldsymbol{\Sigma})$. For each $\boldsymbol{\beta}_s^{(k)}$ we compute the linear predictor $\boldsymbol{\eta}_s^{(k)} = \alpha + \mathbf{X}^\top \boldsymbol{\beta}_s^{(k)}$ and the risk probability vector $\boldsymbol{\pi}_s^{(k)} = \frac{\exp(\boldsymbol{\eta}_s^{(k)})}{1+\exp(\boldsymbol{\eta}_s^{(k)})}$ and finally generate binary response vectors $\mathbf{y}_s^{(k)} \sim \text{Bin}(\boldsymbol{\pi}_s^{(k)}, 1)$. To avoid problems with complete separation, the intercept α has been chosen such that the proportion of events ($y = 1$) is close to 0.5. The simulated data $(\mathbf{y}_s^{(k)}, \mathbf{X})$ are now analysed with R-INLA using a $\text{N}(\boldsymbol{\nu}_{\text{SG}}, g\,\boldsymbol{\Sigma})$ prior for $\boldsymbol{\beta}$ and the following priors on g: Fixed $g = 1$, Hyper-g, horseshoe and Strawderman-Berger.

In total, $27 \times 1000 \times 4 = 108\,000$ calls of R-INLA have been made. Except for fixed $g = 1$, the posterior median of g has been computed and averaged across $K = 1000$ analyses for each ϵ_s. This is shown in the top row of Figure 3. One can see how the different approaches react to prior misspecification with increasing estimates of g for increasing $|\epsilon_s|$. There are only minor differences between the different approaches with a slight bias towards $g > 1$ of the Strawderman prior in the case of no misspecifcation ($\epsilon_s = 0$).

The next row in Figure 3 gives the root mean squared error (RMSE)

$$\sqrt{\frac{1}{K}\sum_{k=1}^{1000}\left\{\mathsf{E}(\beta_j \mid \mathbf{y}_s^{(k)}) - (\nu_j + \epsilon_s)\right\}^2}$$

between the posterior mean of β_j and the true underlying mean $\nu_j + \epsilon_s$, here ν_j denotes the j-th component of the prior mean vector $\boldsymbol{\nu}_{\text{SG}}$. Shown are the results for two covariates, `nullip` with a balanced proportion of 49% "cases" and `hydram` with only 0.3% cases. The third row gives the corresponding mean posterior standard deviation (MPSD) of the two covariates. We show only results for the hyper-g approach, since the other two priors gave virtually identical results. It is interesting to see how the hyper-g approach reacts

to model misspecifcation with much lower RMSE and larger MPSD in the case of model misspecification. As one would expect, differences between hyper-g and fixed-g (both in terms of RMSE and MPSD) increase with increasing amount of misspecifcation.

Finally, the last row in Figure 3 gives the coverage of equi-tailed 95% credible intervals for the components of $\boldsymbol{\beta}$, averaged across all 14 covariates. Whereas the hyper-g and the horseshoe prior (Strawderman gives very similar results) have coverage very close to the nominal 95% level, the empirical coverage of the fixed-g analysis drops quickly to values of 85% and below.

3.4 *Simulation study II with scaled covariance matrix*

In a second simulation study, we have investigated the effect of misspecification of the prior covariance matrix. The study has been conducted as in Section 3.4, with the only difference that $\boldsymbol{\beta}$ is now generated from a $\mathrm{N}(\boldsymbol{\nu}_{\mathrm{SG}}, \delta_s \boldsymbol{\Sigma})$ distribution where the $S = 31$ components of $\boldsymbol{\delta} = (1/20, \ldots, 1, \ldots, 20)$, equally-spaced on the log-scale, quantify the amount of prior misspecification.

The results are shown in Figure 4. The adaptive approaches react to prior misspecification with smaller values of the posterior medians of g for small values of δ_s and *vice versa*. The RMSE of the fixed-g approach increase dramatically for larger δ_s, whereas the increase of the hyper-g approach is only moderate. Of note, the MPSD of the hyper-g is now smaller than for the fixed-g approach for small of δ_s. The coverage of the 95% credible intervals is again very close to the nominal level for the hyper-g approach (Strawderman again not shown, since visually indistinguishable). The horseshoe prior gives similar results, but with coverage slightly too low for small values of δ_s. The fixed-g approach has coverage too high for small values of δ_s and coverage too low for larger values of δ_s.

4. Discussion

We have proposed a novel approach to update the weight of the prior distribution in the light of the current data. We have focused on the common scenario where the prior distribution for the regression coefficients is multivariate normal. Adaptive prior weighting is achieved by introducing an unknown multiplicative scalar g for the prior covariance matrix. A hyperprior for g allows to adaptively estimate the weight of the prior distribution in the light of the current data. The application showed that the hyper-g prior allows for both up- or down-weighting of the prior distribution. Another example with a correlated prior on the regression coefficients is given in Supplementary material.

Prior information on regression coefficients from historical data can often be assumed to be normal due to the approximate normality of the posterior distribution, *e. g*. Bernardo and Smith (2000). A normal prior distribution is therefore a natural choice. The explicit incorporation of a prior weight parameter in our approach can be used to inform researchers on the appropriateness of the original prior being used. The simulation study has shown that the posterior distribution of g informs appropriately about possible misspecification of the prior distribution.

However, if the interest is primarily in the regression coefficients, an alternative way to interpret a hyperprior on the inverse prior weight parameter g is to consider the implied marginal prior distribution on the regression coefficients, a scale mixture of normals (West, 1987). The prior weight $w = 1/g$ is then considered a nuisance parameter and its posterior distribution is only of secondary interest. For example, usage of an inverse gamma hyperprior leads to a "robust" Student t rather than a normal prior distribution (Zellner and Siow, 1980) for the regression coefficients. As a special case one obtains a Cauchy prior as proposed in Gelman et al. (2008) for logistic regression. From that perspective, our approach can be viewed as replacing a normal prior on the regression

coefficients with a "robustified" scale mixture of normals prior. In contrast to Student t or Cauchy priors, the hyper-g prior on g allows for a symmetric treatment of the weight parameter and can be viewed as a regularized version of empirical Bayes, thus balancing prior-data agreement and disagreement. However, the implied marginal distribution of the regression coefficients does not have a standard form (Liang et al., 2008).

As an extension of our approach we have introduced several independent weight parameters for blocks of regression coefficients in the application described at the end of Section 3.1. In the limit, even every regression coefficient can have its own weight parameter as long as the prior covariance matrix is diagonal. The advantage of the implementation in INLA is that the proposed methodology can easily be applied in more complex models, *e. g*. generalized linear mixed models as in our second application described in Supplementary Material. In future work we also plan to compare the sensitivity of the posterior of the regression coefficients (with and without adaptive prior weighting) with respect to mean and covariance matrix of the normal prior (Roos and Held, 2011; Roos et al., 2015).

Supplementary Material

Supplementary Material describes the implementation of the proposed methodology with INLA in detail and an application longitudinal data using a generalized linear mixed model.

Acknowledgments

R-code by Daniel Sabanés Bové, written for the analysis described in Held et al. (2012), was helpful to implement the approach proposed in this paper.

References

Albert, A. and Anderson, J. A. (1984). On the existence of maximum-likelihood estimates in logistic regression models. *Biometrika* **71,** 1–10.

Berger, J. (1980). A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *Annals of Statistics* **8,** 716–761.

Bernardo, J. M. and Smith, A. F. M. (2000). *Bayesian Theory*. Wiley Series in Probability and Statistics. John Wiley & Sons, Chichester.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A (General)* **143,** 383–430.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97,** 465–480.

Copas, J. B. (1983). Regression, prediction and shrinkage. *Journal of the Royal Statistical Society. Series B (Methodological)* **45,** 311–354.

Copas, J. B. (1997). Using regression models for prediction: shrinkage and regression to the mean. *Statistical Methods in Medical Research* **6,** 167–183.

Cui, W. and George, E. I. (2008). Empirical Bayes vs. fully Bayes variable selection. *Journal of Statistical Planning and Inference* **138,** 888–900.

Denison, D. G. T., Holmes, C. C., Mallick, B. K., and Smith, A. F. M. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. Wiley Series in Probability and Statistics. Wiley, Chichester.

Duan, Y., Ye, K., and Smith, E. P. (2006). Evaluating water quality using power priors to incorporate historical information. *Environmetrics* **17,** 95–106.

Evans, M. and Jang, G. H. (2010). Invariant *P*-values for model checking. *The Annals of Statistics* **38,** 512–525.

Evans, M. and Jang, G. H. (2011a). A limit result for the prior predictive applied to

checking for prior-data conflict. *Statistics & Probability Letters* **81,** 1034–1038.

Evans, M. and Jang, G. H. (2011b). Weak informativity and the information in one prior relative to another. *Statistical Science* **26,** 423–439.

Evans, M. and Moshonov, H. (2006). Checking for prior-data conflict. *Bayesian Analysis* **1,** 893–914.

Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics* **13,** 342–368.

Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing* **7,** 57–68.

Gelman, A., Jakulin, A., Grazia, M. P., and Yu-Sung, S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics* **2,** 1360–1383.

Greenland, S. (2006). Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *International Journal of Epidemiology* **35,** 765–775.

Greenland, S. (2007a). Bayesian perspectives for epidemiological research. II. Regression analysis. *International Journal of Epidemiology* **36,** 195–202.

Greenland, S. (2007b). Prior data for non-normal priors. *Statistics in Medicine* **26,** 3578–3590.

Greenland, S. (2009). Bayesian perspectives for epidemiologic research: III. Bias analysis via missing-data methods. *International Journal of Epidemiology* **38,** 1662–1673.

Greenland, S. and Mansournia, M. A. (2015). Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Statistics in Medicine* **34,** 3133–3143.

Held, L., Sabanés Bové, D., and Gravestock, I. (2015). Approximate Bayesian model

selection with the deviance statistic. *Statistical Science* **30,** 242–257.

Held, U., Sabanés Bové, D., Steurer, J., and Held, L. (2012). Validating and updating a risk model for pneumonia - a case study. *BMC Medical Research Methodology* **12,** 99.

Hoerl, A. E., Kennard, R. W., and Baldwin, K. F. (1975). Ridge regression: Some simulations. *Communications in Statistics: Theory & Methods* **4,** 105–124.

Ibrahim, J. G. and Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Science* **15,** 46–60.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90,** 773–795.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association* **103,** 410–423.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B (Methodological)* **34,** 1–41.

Marin, J.-M. and Robert, C. P. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer texts in Statistics. Springer, New York.

Martins, T. G., Simpson, D., Lindgren, F., and Rue, H. (2013). Bayesian computing with INLA: new features. *Computational Statistics and Data Analysis* **67,** 68–83.

Miettinen, O. S., Flegel, K. M., and Steurer, J. (2008). Clinical diagnosis of pneumonia, typical of experts. *Journal of Evaluation in Clinical Practice* **14,** 343–350.

Neuenschwander, B., Branson, M., and Spiegelhalter, D. J. (2009). A note on the power prior. *Statistics in Medicine* **28,** 3562–3566.

Neutra, R. R., Fienberg, S. E., Greenland, S., and Friedman, E. A. (1978). Effect of fetal monitoring on neonatal death rates. *New England Journal of Medicine* **299,** 324–326.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J.,

Oakley, J. E., and Rakow, T. (2006). *Uncertain Judgements; Eliciting Experts' Probabilities*. Wiley, Chichester.

Roos, M. and Held, L. (2011). Sensitivity analysis in Bayesian generalized linear mixed models for binary data. *Bayesian Analysis* **6,** 259–278.

Roos, M., Martins, T. G., Held, L., and Rue, H. (2015). Sensitivity analysis for Bayesian hierarchical models. *Bayesian Analysis* **10,** 321–349.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society - Series B* **71,** 319–392.

Sabanés Bové, D. and Held, L. (2011). Hyper-g priors for generalized linear models. *Bayesian Analysis* **6,** 387–410.

Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Spiegelhalter, D., and Neuenschwander, B. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics* **70,** 1023–1032.

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Wiley, New York.

Sullivan, S. G. and Greenland, S. (2013). Bayesian regression in SAS software. *International Journal of Epidemiology* **42,** 308–317.

West, M. (1987). On scale mixtures of normal distributions. *Biometrika* **74,** 646–648.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In Goel, P. K. and Zellner, A., editors, *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, volume 6 of *Studies in Bayesian*

Econometrics and Statistics, chapter 5, pages 233–243. North-Holland, Amsterdam.

Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M., editors, *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia*, pages 585–603, Valencia. University of Valencia Press.

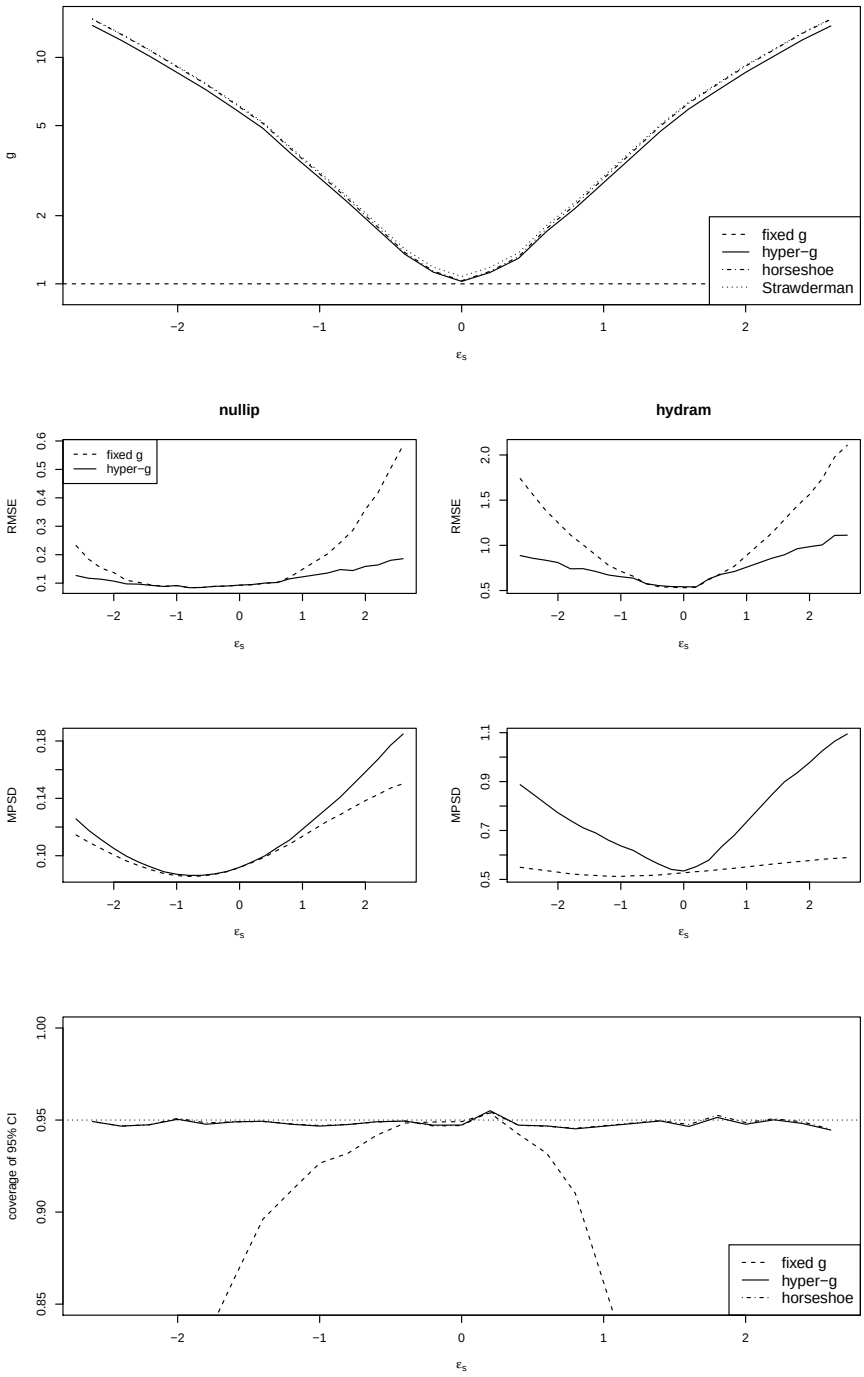

Figure 3: Simulation I: Mean posterior median estimates of g (top row), root mean squared differences (RMSE) and mean posterior standard deviation (MPSD) for two explanatory variables (`nullip` and `hydram`) (middle rows) and average coverage of 95% credible intervals across all explanatory variables (bottom row).

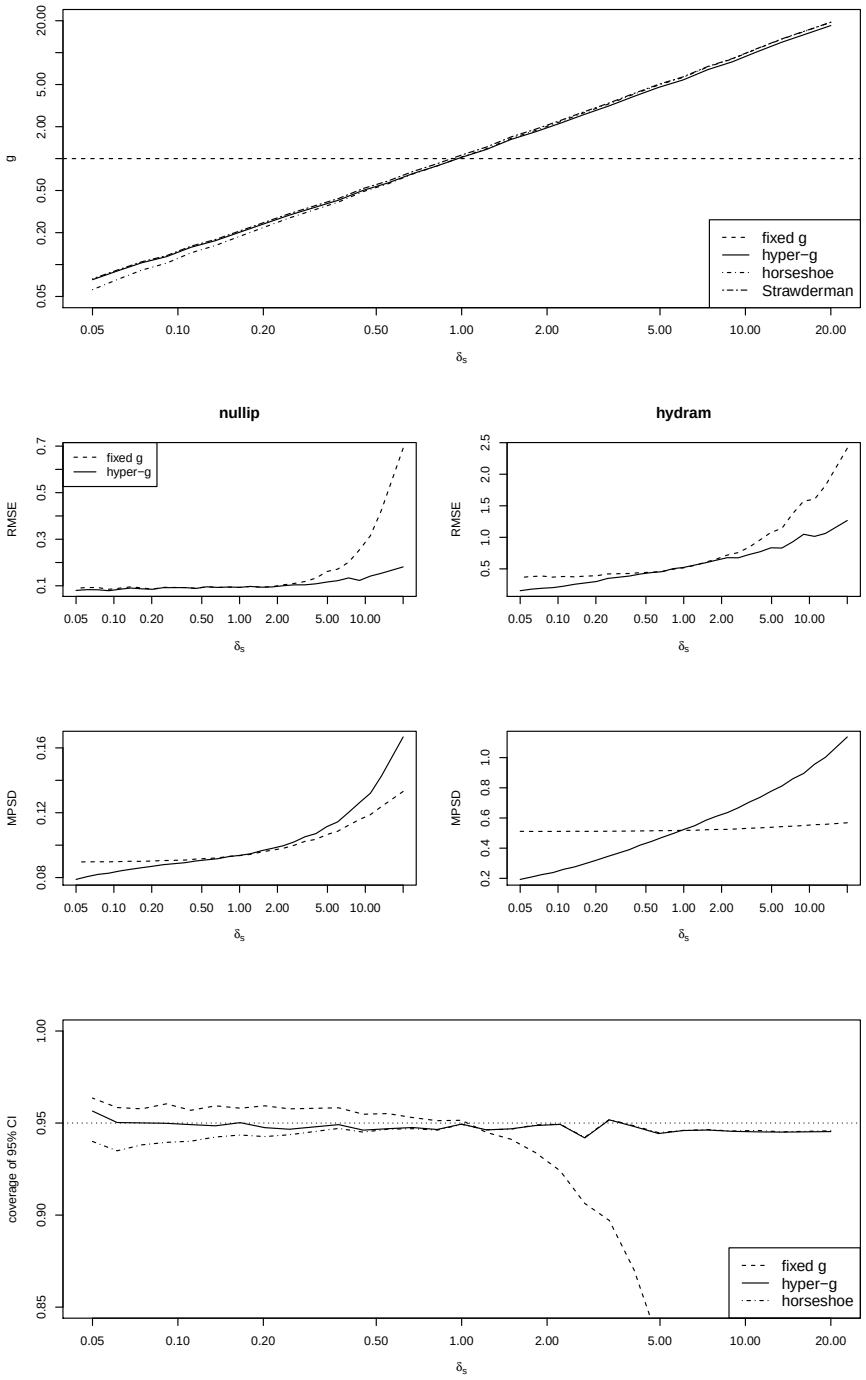

Figure 4: Simulation II: Mean posterior median estimates of g (top row), root mean squared differences (RMSE) and mean posterior standard deviation (MPSD) for two explanatory variables (`nullip` and `hydram`) (middle rows) and average coverage of 95% credible intervals across all explanatory variables (bottom row).

Supplementary material for

"Adaptive prior weighting in generalized regression"

Leonhard Held and Rafael Sauter

Department of Biostatistics

Epidemiology, Biostatistics and Prevention Institute

University of Zurich

Hirschengraben 84, 8001 Zurich, Switzerland

Email: leonhard.held@uzh.ch, rafael.sauter@uzh.ch

This supplementary material guides through technical details of the applications presented in Section 3 in the main text. We discuss the implementation of a multivariate normal prior

$$\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{\nu}, g\boldsymbol{\Sigma}) \tag{1}$$

on the regression coefficients $\boldsymbol{\beta}$ in a generalized regression model with integrated nested Laplace approximations (INLA) (Rue et al., 2009). INLA is implemented, freely available and comes with a `R` user interface (`r-inla`) which can be installed by the following `R` command

```
install.packages("INLA", repos="http://www.math.ntnu.no/inla/R/stable").
```

Information about the installed version can be retrieved by typing

```
inla.version().
```

For this document and the applications in the main text we used the `r-inla` version built on 2014-07-04. In Section 1 we introduce how a generalized linear model (GLM) is defined with `r-inla`. Section 2 illustrates the implementation of a user defined prior in `r-inla` and Section 3 provides the code to reproduce the results of the applications presented in the main text as well as an additional example of a clinical trials with a longitudinal data structure.

1. Generalized regression with INLA

A generalized linear model (GLM) is described by a likelihood for the observations y_i with $i = 1, \ldots, n$ and a linear predictor $\eta_i = \alpha + \boldsymbol{x}_i^\top \boldsymbol{\beta}$, where $\boldsymbol{\beta}$ has length d. A Bayesian analysis requires prior distributions on all unknown parameters. Different likelihoods and priors are implemented in `r-inla`. An overview, descriptions and implementation

details can be found on `http://www.r-inla.org/`. In Section 1.1 we describe the implementation of the multivariate prior (1) in `r-inla`.

1.1 *Multivariate normal priors in INLA*

The model analyzed by the `inla` function is defined by a formula, similar to a `glm` formula (see `?inla` and `?formula`). The function `inla()` computes the marginal posterior distribution for all model parameters, defined in the `formula` argument and generates an `inla`-object. In `r-inla` a latent model is defined by calling `f(..., model = )`, where `...` must be replaced by a variable name describing the dependence structure in the latent model. Details about the available models can be found by typing `?f`.

To implement the prior (1) with mean $\boldsymbol{\nu} = \mathbf{0}$, we use the latent model `"generic0"`. The `"generic0"` latent model allows to use a generic multivariate normal prior with precision matrix $\mathbf{Q} = w\,\mathbf{C}$. A prior distribution on the precision hyperparameter w can be defined. In the context of the main text we use $\mathbf{C} = \boldsymbol{\Sigma}^{-1}$ and $w = 1/g$. The following `R`-code shows a sample model-formula, which uses the `generic0`-field for $d = 2$ covariates:

```
#define inla-model:
inlaFormula <-    1 + f(indexName1,
                        varName1,
                        model = "generic0",
                        Cmatrix = priorPrecision,
                        initial = log(0),
                        fixed = FALSE,
                        hyper = list(theta = list(
                                                  prior = "loggamma",
                                                  param = c(a, b)
```

```
                                        ) ) )
            + f (indexName2, varName2, copy = "indexName1")
```

As in the `glm`-models `"-1"` would remove the global intercept. Usually, if we assume a separate improper prior on the intercept (α), we keep the intercept, in the same way as in the code above. In the GLM application `indexName`p refers to the p^{th} covariate and consists only of integers equal to p defining an index variable among d different covariates. `varName`p gives the covariate name, referring to the corresponding variable and is used as weight in the latent field. The `Cmatrix` argument defines the precision matrix $\mathbf{C} = \mathbf{\Sigma}^{-1}$. Thus `priorPrecision` is a covariance matrix of dimensions $d \times d$. The argument `initial` is used to set a starting value for w and `fixed` indicates whether w should be fixed at its starting value. The argument `hyper` defines the prior distribution for w which in this case is a log-Gamma distribution with parameter `a` and `b` on the log-precision $\theta = \log(w)$, which is the `r-inla` internal scale for the hyperparameter w. The settings for the `hyper` argument are ignored if `fixed = TRUE`.

If the GLM includes $d > 1$ covariates then we need to call additional $(d-1)$ latent fields using the `copy` argument defined as `f(..., copy="indexName1")`. This `copy` feature is used to define the latent field correctly if more than one covariate is included and weights each component of the latent field, as described in more detail in Martins et al. (2013, Section 4.3). Thus, the object `inlaFormula` defined above, includes $d = 2$ covariates (`varName1` and `varName2`). The following `R`-code calls the `inla` function and computes the model defined above:

```
#inla-call:
result <-  inla(formula = inlaFormula,
                data = data,
```

```
family = "binomial",
family = paste(inla.family),
offset = Offs,
control.fixed=list(mean = PriorMeanIntercept,
                   prec = PriorPrecIntercept
                   ))
```

The likelihood of the model is defined by the argument `family` and a data-frame or a list with the requested variable names must be provided in the `data` argument. The argument `control.fixed` allows to set prior distribution parameters for fixed effects and the intercept. We assume that all covariates are in the latent field except the global intercept for which we assume a prior mean defined by the numeric value `PriorMeanIntercept` and a prior precision defined by the numeric value `PriorPrecIntercept`.

As the `generic0` model assumes a zero mean prior, we need to define suitable offsets $o_i = \boldsymbol{x}_i^\top \boldsymbol{\nu}$ to allow for a non-zero prior mean $\boldsymbol{\nu}$. Adding the offset implies that $\eta_i = \alpha + \boldsymbol{x}_i^\top \boldsymbol{\nu} + \boldsymbol{x}_i^\top \widetilde{\boldsymbol{\beta}}$. If one is interested to use different priors on separate blocks of covariates, as discussed in Section 3.1 in the main text, one needs to repeatedly define separate `generic0` latent-fields for each block, which is explained in more detail in Section 3.2. In Section 1.2 we introduce the function `pwGLM` which is a wrapper function calling `r-inla` and which facilitates the implementation of the multivariate normal prior defined in (1).

1.2 *Adaptive prior weighting with the `pwGLM` function*

In order to appropriately define the latent field, in dependency of the covariates $\boldsymbol{\beta}$ included in the GLM, the prior means $\boldsymbol{\nu}$ and prior covariance $\boldsymbol{\Sigma}$, we wrote the `R`-function `pwGLM` which generates the corresponding model. The function is available in the supplementary `R`-package `pwGLMinla`. This function was adapted and extended from an

earlier implementation used by Held et al. (2012). The function `pwGLM` takes the following arguments (default settings are given after `" = "`):

- `response`: the response vector in the regression model.
- `data`: a data frame with the covariates in the regression model.
- `priorMean`: a vector with the prior mean ν of length d, the same as the number of covariates.
- `priorPrecision`: a precision matrix with dimension $d \times d$. If separate priors for k covariate blocks is used (*i.e.* `sepG` is a vector) then `priorPrecision` is used for every covariate block (only possible if each of k blocks is of equal dimension) or `priorPrecision` needs to be a list of length k with each list element containing a precision matrix of suitable size fitting to the corresponding covariate block.
- `priorG = "loggamma"`: defines the hyper-g prior on $\log(1/g)$, either the name of an INLA built-in prior, a table or expression. If separate priors for k covariate blocks is used (*i.e.* `sepG` is a vector) then `priorG` is applied to every of the k covariate blocks or needs to be a list of length k defining possibly different priors for each covariate block.
- `HyPar = NULL`: parameters for the prior-distribution of the hyper-g prior. If `priorG` is a table or expression or if `fixedG=TRUE` then `HyPar` will be ignored. If separate priors for k covariate blocks is used (*i.e.* `sepG` is a vector) then `HyPar` is applied to every of the k covariate blocks or needs to be a list of length k defining possibly different parameters of the prior distributions for each covariate block.
- `initialG = 0`: starting value for the hyper-g prior. If `priorG` is a table or expression then `initialG` will be ignored. If separate priors for k covariate blocks is used (*i.e.* `sepG` is a vector) then `initialG` is applied to every of the k covariate blocks or needs to be a vector of length k defining possibly different starting values for the hyperparameters for each covariate block.

- `fixedG = FALSE`: should hyper-g prior be fixed at `initialG`. If `fixedG = TRUE` then the prior distribution on hyper-g (`HyPar`) will be ignored. If separate priors for k covariate blocks is used (*i. e.* `sepG` is a vector) then `fixedG` is applied to every of the k covariate blocks or needs to be a vector of length k defining possibly different fixed hyperparameters for each covariate block.

- `sepG = 1`: defines separate covariate blocks for which a separate `generic0` latent-field will be implemented and thus also a separate hyper-g prior will be used. If `sepG = 1` (default) then only one covariate block is used. If `sepG` is a vector of length d, the same length as the number of covariates, with k different entries, then each covariate will be assigned to the corresponding covariate block. The k entries in `sepG` should be integers $1, 2, \ldots, k$.

  WARNING: The order of the covariates (*e. g.* in `formula`) must correspond to the index in `sepG`. If `sepG` is a vector it will affect the meaning of other options (`priorPrecision`, `priorG`, `HyPar`, `fixedG` and `initialG`).

- `intsep = TRUE`: should the intercept be treated separately without any hyper-g prior. If `intsep = FALSE`, a possible intercept must be included like the other covariates and the precision matrix must be extended correspondingly.

- `PriorMeanIntercept = 0`: prior mean for intercept if `intsep = TRUE`.

- `PriorPrecIntercept = 0`: prior precision for intercept if `intsep = TRUE`.

- `FormExt = NULL`: an extension of the model formula which is not covered by $\boldsymbol{\beta}$ (*e. g.* an additional latent field for patient-specific random intercepts in the case of a generalized linear mixed model).

- `AddDat = NULL`: additional data only used in the additional latent field and only requested if `FormExt!=NULL`.

- `inla.family = "binomial"`: the GLM-family name defining the likelihood, default is a binomial distribution, see `www.r-inla.org/models/likelihoods`.
- `verbose = TRUE`: prints the INLA-formula, is not the INLA-verbose argument.
- `inla.strat = list(strategy = "simplified.laplace", int.strategy = "grid", dz = 0.3)`: control variables in `inla` which are defined differently than the `inla` defaults, see `?control.inla`.
- `updateHyper = TRUE`: should the marginal posterior of the hyperparameter be updated *i.e.* computed more precisely requiring more computing time, see `?inla.hyperpar`.

A call to `pwGLM` returns a list summarizing the results with the following named elements containing the following values:

- `coefNames`: the names of the coefficients included in the model.
- `inlaResult`: the `inla` object returned by the `inla` call. Essentially all the information in the other elements of the list returned by `pwGLM` can also be found here and are derived form this object. In case of any doubts one should always check with the content in the `inlaResult` list element.
- `betaMedian`: the median of the marginal posterior distributions for the included covariates $\boldsymbol{\beta}$.
- `betaMean`: the mean of the marginal posterior distributions for the included covariates $\boldsymbol{\beta}$.
- `betaFixed`: the prior-mean $\boldsymbol{\nu}$, which is used to compute the offset $o_i = \boldsymbol{x}_i^\top \boldsymbol{\nu}$.
- `gQuantiles`: the quantiles (2.5%, 50% and 97.5%) of the marginal posterior distribution of g (or several g's if `sepG` is a vector).

The function `pwGLM` is primarily used to construct the INLA model formula which can be evaluated by the `inla`-function. Within `pwGLM` the function `inla` from the `r-inla` package

is called. Additional arguments to the function inla may be assigned in the `pwGLM` call which are then passed to `inla`.

Examples in Section 3 show the detailed use of `pwGLM`. The function `pwGLM` together with the datasets used for the applications in Section 3 in the main text is made readily available as `R`-package `pwGLMinla` in order to facilitate the reproduction of the presented results.

2. Defining different prior distributions for g

In Section 2.1 we show how to implement the hyper-g prior with $a = 4$ such that it can be used in `inla` and respectively by `pwGLM`. The hyper-g prior is used in Section 3 and for the same applications discussed in the main text. The hyper-g prior with $a = 4$ corresponds to a standard uniform distribution on the shrinkage factor $t = \frac{g}{1+g}$ which is equivalent to Beta distribution $t \sim \text{Be}(1, 1)$. Alternatively to a hyper-g prior one could be interested in specifying any other Beta distribution on t. The implementation and necessary transformations for a $t \sim \text{Be}(a_1, a_2)$ distribution is illustrated in Section 2.2.

The package `r-inla` allows to use certain pre-specified priors as well as user defined priors. In this section we demonstrate the usage of a user defined prior. As `r-inla` internally uses often a log-transformation of the parameter and because one occasionally wishes to switch between a prior specification on g or on the shrinkage factor $t = \frac{g}{1+g}$ one needs to apply the change-of-variables formula (see Held and Sabanés Bové, 2014, Appendix A.2.3) to transform the prior densities: if $f_X(x)$ is a probability density function of X one can compute the probability density function $f_Y(y)$ of the transformed variable $Y = f(X)$ by

$$\begin{aligned} f_Y(y) =& f_X\{f^{-1}(y)\} \left| \frac{df^{-1}(y)}{dy} \right| \\ =& f_X(x) \left| \frac{df(x)}{dx} \right|^{-1}. \end{aligned}$$

2.1 *Hyper-g prior*

In `r-inla` it is possible to implement a user-defined prior for the hyperparameters (g in in our case). This can be done either by defining a function by using the `muparser` library or by defining a table with the prior evaluated on a suitable grid (see `http://www.r-inla.org/models/priors`). We choose the tabulated version for the implementation presented below.

Change of variables for hyper-g. We want to use the hyper g-prior $f_g(g) = \frac{a-2}{2}(1+g)^{-a/2}$ with $a = 4$ such that $f_g(g) = (1+g)^{-2}$. The latent model `generic0` in `r-inla` is parameterized with precision $\mathbf{Q} = w\,\mathbf{C}$ where $w = 1/g$ and $\mathbf{C} = \mathbf{\Sigma}^{-1}$. Internally `r-inla` uses $\theta = \log(w)$. Thus we need to transform the prior density appropriately to θ using the change-of-variables formula: applied here for the transformation $\theta = f(g) = \log(1/g)$ and $f^{-1}(\theta) = \exp(1/\theta) = \exp(-\theta)$.

The absolute value of the derivative is $\left|\frac{df^{-1}(\theta)}{d\theta}\right| = |-\exp(-\theta)| = \exp(-\theta)$ and thus the probability density function of θ is

$$\begin{aligned} f_\theta(\theta) =& f_g\{f^{-1}(\theta)\}\left|\frac{df^{-1}(\theta)}{d\theta}\right| \\ =& (1+\exp(-\theta))^{-2}\cdot\exp(-\theta). \end{aligned}$$

In `r-inla` the log-density is required: $\log(f_\theta(\theta)) = \log\{(1+\exp(-\theta))^{-2}\cdot\exp(-\theta)\}$. This is the input prior-distribution needed for `r-inla` corresponding to a hyper-g prior on g.

R-function for hyper-g prior:. Function implemented in R:

```
hypergprior <- function(t){
                            densg <- ((1+exp(-t))^-2)*exp(-t)
                            logdensg <- log(densg)
                            return(logdensg)
                           }
```

Generate a table which can be used by `inla`*:.* The `hypergprior`-function must be evaluated on a suitable grid and stored in an object which can be called by `inla`.

```
#define a suitable grid:
lprec <-  seq(-100, 100, len=20000)

#evaluate prior:
prior.table <- paste(
                     c("table:",
                       cbind(lprec, sapply(lprec,FUN=hypergprior))
                      ), sep = "", collapse = " ")
```

In Section 3.1, 3.2 and 3.3 the hyper-g prior will be applied by calling the object `prior.table`.

2.2 *Beta distribution on t*

The hyper g-prior implies a Beta distribution on the shrinkage factor of the form $t \sim \text{Be}(1, a/2-1)$. So far we looked at the hyper-g prior with $a = 4$, assuming a standard uniform on t. But one could define alternative Beta priors on t. Beta-priors on t which could be of interest is *e.g.* the case where $(a < 4)$. With $a = 3$ we get the Strawderman-Berger prior $t \sim \text{Be}(1, 1/2)$ or the Horseshoe prior for with $t \sim \text{Be}(1/2, 1/2)$.

In order to implement a Beta prior on t we must again apply the transformation of variables. As the `r-inla` internal scale is again on $\theta = \log(w) = \log(1/g)$ the transformation function to the shrinkage factor is $\theta = f(g) = \log(1/g) = \log\left(\frac{1-t}{t}\right)$, the inverse function is $f^{-1}(\theta) = \frac{1}{(1+\exp(\theta))}$ and the absolute value of the derivative $\left|\frac{df^{-1}(\theta)}{d\theta}\right| = \left|\frac{\exp(\theta)}{(1+\exp(\theta))^2}\right|$. For $t \sim \text{Be}(a_1, a_2)$ the probability density function is $f_t(t) = B(a_1, a_2)^{-1} t^{a_1-1}(1-t)^{a_2-1}$ (see Held and Sabanés Bové, 2014, Appendix A.5.2). Applying the change-of-variables

formula we get the probability density function for θ to be

$$\begin{aligned} f_\theta(\theta) =& f_t\{f^{-1}(\theta)\}\left|\frac{df^{-1}(\theta)}{d\theta}\right| \\ =& B(a_1,a_2)^{-1}\left(\frac{1}{1+\exp(\theta)}\right)^{a_1-1}\left(1-\frac{1}{(1+\exp(\theta)}\right)^{a_2-1}\frac{\exp(\theta)}{(1+\exp(\theta))^2}. \end{aligned}$$

In `r-inla` again the log-density is required: $\log(f_\theta(\theta))$ which is the input prior-distribution needed for `r-inla` for defining the prior on $\theta = \log(1/g)$ corresponding to a Beta-prior on the shrinkage factor t.

R-function for Beta prior on t:. Function implemented in R:

```
betatprior <- function(t, a1=NULL, a2=NULL){
  tinv <- 1/(1+exp(t))
  densbeta <- dbeta(tinv, shape1=a1, shape2=a2)*exp(t)/(1+exp(t))^2
  logdensbeta <- log(densbeta)
  return(logdensbeta)
                          }
```

Generate a table which can be used by `inla`:. The `betatprior`-function must be evaluated on a suitable grid and stored in an object which can be called by `inla`.

```
#define a suitable grid:
lprec <-  seq(-100, 100, len=20000)     ##CHANGE this LINE if INLA crashes!


#evaluate prior:
prior.table.t <- paste(
                     c("table:",
                       cbind(lprec, sapply(lprec,FUN=betatprior,
                                           a1=0.5, a2=0.5))),
                       sep = "", collapse = " ")
```

3. Applications

In this section we demonstrate the application of adaptive prior weighting by using `pwGLM` for the logistic regression example discussed by Sullivan and Greenland (2013) in Section 3.1. The application of three separate g-priors on covariate blocks for this dataset is demonstrated in Section 3.2 and the application of prior weighting to a binary response generalized linear mixed model for longitudinal data about the comparison of two treatments against a toenail infection in Section 3.3. In order to be able to run the examples from Section 3.1, 3.2 and 3.3 make sure to install the accompanying R-package `pwGLMinla` first (see `?install.packages`).

3.1 *Adaptive prior weighting in a logistic regression model*

```
#load library
library(pwGLMinla)

#load data
data(logregdat)

#rescale dyslab:
logregdat$dyslab=logregdat$dyslab/3

#Design matrix w/o intercept:
logregdatX<- subset(logregdat, select=-death)
Ntrials <- rep(1, nrow(logregdatX))
```

```
#define priorMeanG:

priorMeanG <- log(c(2,2,2,4,2,1,4,2,2,2,4,2,2,4))


#define priorPrecisionG:

priorPrecisionG <- diag(rep(2, ncol(logregdatX)))


#call pwGLM:

resultLog <- pwGLM( response = subset(logregdat, select=death),
                    data = logregdatX,
                    priorG = prior.table,
                    priorMean = priorMeanG,
                    priorPrecision = priorPrecisionG,
                    verbose = TRUE,
                    inla.family = "binomial",
                    Ntrials = Ntrials, # number of trials for binomial
                    intsep = TRUE)


#Plot the marginal posterior of g:

#1. extract the marginal posterior from INLA-object:

marg.g.inv <- resultLog$inlaResult$marginals.hyperpar[[1]]


#2. use inla.tmarginal to transform from w=1/g to g:

marg.g <- inla.tmarginal(function(x) 1/x, marg.g.inv, method="linear")
```

```
#3. Plot:
plot(marg.g[,"x"], marg.g[,"y"], type="l",
     xlim=c(0,2), ylim=c(0,1), xlab="g", ylab="density")

#Plot the marginal posterior of hydram:
#1. Extract the prior Mean of hydram (covariate with index 7):
m1 <- resultLog$betaFixed[7]

#2. Extract the marginal posterior of hydram from the latent field:
beta <- resultLog$inlaResult$marginals.random$ihydram$index.7

#3. Add prior mean
beta[,1] <- beta[,1] + m1

#4. Plot
plot(beta, type = "l",
     xlab = "beta", ylab = "density")
```

3.2 *Adaptive prior weighting with separate hyper-g priors*

Instead of using one single g for $\boldsymbol{\beta}$ one can use several g-priors on covariate blocks, as mentioned in Section 3.1 in the main text. Here we show how the case of three separate g's can be implemented with the `pwGLM(...)` function.

First we define again the prior mean, which is set here to zero and the prior covariance matrix. As we want to use three different g-prior on three different covariate blocks we need to define three separate covariance matrix of adequate size and join them in a

single list. The first block addresses nine covariates for which in the above example a prior mean equal to $\log(2)$ was assigned. The second covariate block of size four covers the covariates with previous prior mean equal to $\log(4)$ and the last g addresses a single variable (`abort`) with previous prior mean equal to $\log(1)$.

```
#Prior mean set to zero:
priorMeanZero <- rep(0, length(priorMeanG) )

#Three separate prior (diagnoal) covariance matrix:
priorlogsepg1 <- diag(9)*2
priorlogsepg2 <- diag(4)*2
priorlogsepg3 <- 2

#Join the three prior covariance matrix as list:
priorlistLOG <- list(
                priorlogsepg1,
                priorlogsepg2,
                priorlogsepg3)
```

Then we apply `pwGLM` which calls `r-inla`. The argument `sepG` defines which variables are addressed by which covariate block, where the order corresponds to the order of the variables in the data object `logregdatX`.

```
#call pwGLM:
resultLogSepG <- pwGLM(response=subset(logregdat, select=death),
                    data=logregdatX,
                    priorG=prior.table,
```

```
                priorMean=priorMeanZero,
                priorPrecision=priorlistLOG,
                fixedG=FALSE,
                sepG=c(1,1,1,2,1,3,2,1,1,1,2,1,1,2),
                verbose=TRUE,
                Ntrials=Ntrials,  # number of trials for binomial
                intsep=TRUE,
                updateHyper=FALSE,
                inla.strat=list(strategy = "simplified.laplace",
                                int.strategy = "grid", dz=0.8),
                )

#The three separate g are stored in object resultLogSepG:
gsep <- resultLogSepG$gQuantiles
```

3.3 *Bayesian analysis of binary longitudinal data*

A randomized, double-blinded clinical trial has been conducted to compare a novel oral treatment A for toenail infections against standard therapy (treatment B). Results from this study have been published in De Backer et al. (1998). Data are available on `http://onlinelibrary.wiley.com/journal/10.1111/%28ISSN%291467-9876/homepage/50_3.htm`.

The data, also described in detail by Molenberghs and Verbeke (2005), include 1908 observations from I=294 patients. Patient $i = 1, \ldots, I$ has longitudinal observations y_{ij} at occasions $j = 1, \ldots, n_i$ with associated timepoints t_{ij}. At each visit, patients were evaluated for the degree of onycholysis. The outcome was rated absent, mild, moderate or severe onycholysis and was dichotomized to a binary response with either absent or

mild ($y_{ij} = 0$) or moderate to severe ($y_{ij} = 1$) onycholysis. The first observation was always at baseline ($t_{ij} = 0$) prior to treatment. The follow-up visits were planned to take place after 1, 2, 3, 6, 9 and 12 months since the initial visit. The actually observed time since baseline t_{ij} is somewhat varying from patient to patient and is recorded in months. There are also some missing values, so the dataset is unbalanced.

We present here how the prior weighting for the toenail application can be implemented with `pwGLM`. The toenail dataset in its unedited form is available in the `R`-package `pwGLMinla`. There are some data preparation steps necessary before it can be used: `r-inla` and `pwGLM`.

```
#load data
data(toe)
# Data preparation
#rename treatment:
toe$trt <- toe$Treatment
#rename time variable:
toe$time <- toe$Month
#create time:Treatment interaction variable:
toe$timetrt <- toe$time
toe$timetrt[toe$Treatment==0] <- 0
toe$timectrt <- toe$timec
toe$timectrt[toe$Treatment==0] <- 0
#remove 5 patient-id with only one obs:
idrm <- as.numeric(names(which(table(toe$ID)==1)))
#Five patients with ID 45  48  63  99 377 have only on baseline visit.
#These will be removed from the toe-data.
```

```
toe <- toe[-which(toe$ID%in%idrm), ]
#number of observations:
n <- dim(toe)[1]
#replace id-names for toe as r-inla expects seq(1:n) as id-name:
m <- length(unique(toe$ID)) #number of patient-ID's
for(i in 1:m){
  idi <- unique(toe$ID)[i]
  toe[which(toe$ID==idi), "id1"] <- i
  } #end for loop
#remove duplicate variables:
toe <- toe[, -which(names(toe)%in%c("Treatment","Month", "ID")) ]
#Ntrials argument for r-inla stored in a separate object:
Ntrials <- rep(1, nrow(toe))
```

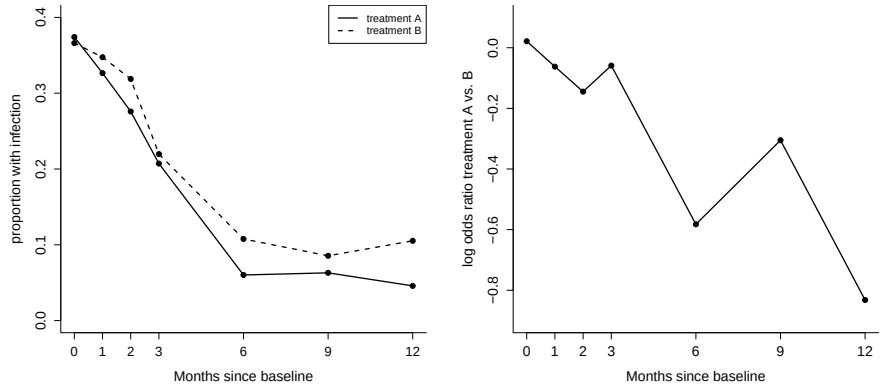

Figure 1: Proportion of toenail infections and empirical log odds ratio for treatment A and B.

The left plot in Figure 1 shows the proportion of patients with a (moderate or severe)

toenail infection receiving treatment A or B at each of the scheduled follow-up visits. The right plot shows the empirical log odds ratio of treatment A versus B for a toenail infection.

A logistic regression model with independent patient-specific random intercepts $b_i \sim \mathrm{N}(0, \sigma_b^2)$, *i. e.* $\mathrm{logit}(\pi_{ij}) = \alpha + \mathbf{x}_{ij}^\top \boldsymbol{\beta} + b_i$, is used to describe the probability of infection $\pi_{ij} = \mathsf{Pr}(y_{ij} = 1)$. The covariate vector $\mathbf{x}_{ij} = (\mathrm{trt}_i, t_{ij}, t_{ij} \times \mathrm{trt}_i)^\top$ includes a binary treatment indicator trt_i (0 for control treatment B, 1 for new treatment A), time measured in month (t_{ij}), and an interaction term of time and treatment ($t_{ij} \times \mathrm{trt}_i$). Note that the main interest is in this interaction parameter $\beta_{\mathrm{time}\times\mathrm{trt}}$, as treatment differences at baseline, represented by the main effect β_{trt}, are expected to be small.

For a Bayesian analysis we follow Fong et al. (2010), who have proposed an inverse gamma prior for the random intercept variance in logistic regression models: $\sigma_b^2 \sim \mathrm{IG}(0.5, 0.0164)$. For the regression coefficients $\boldsymbol{\beta}$ we specify a normal prior distribution with mean $\boldsymbol{\mu}$ and covariance matrix $g\,\boldsymbol{\Sigma}$. The components of $\boldsymbol{\mu}$ are zero except for the effect of time t_{ij} in the control group, where we expect an overall drop of infection prevalence from 0.5 to 0.1 in 12 months. This corresponds to log odds of $\log(1/9)/12 = -0.183$ per month, so $\boldsymbol{\mu} = (0, -0.183, 0)^\top$.

The following code defines the prior mean and prior covariances (and precision) for the fixed effects $\boldsymbol{\beta}$:

```
#Prior mean:
mean.toe <- c(0, log(1/9)/12, 0)


# Correlation between effects:
vec <- c(0,1,2,3,6,9,12)
sel <- length(vec)
```

```
x1 <- c(rep(1,sel), rep(0,7))

x2 <- rep(vec, 2)

x3 <- x1*x2

X <- cbind(x1,x2,x3)

my.cor <- cov2cor(t(X)%*%X)


# function Greenland computes the variance of

# normal prior for limits of 95\% CI credible

# interval on OR scale

greenland <- function(lower, upper, level=0.95, scale=1){

    z <- qnorm((1+level)/2)

    num <- (log(upper)-log(lower))/scale

    my.var <- (num/(2*z))^2

    return(my.var)

}


# treatment effect at baseline is expected to be

# between 0.9 and 10/9 with prob 95%

v1 <- greenland(0.9, 10/9)


# probability of infection in control group is

# expected to change with not more than 0.9 --> 0.1 (or 0.1 --> 0.9)

# with prob 95% in 12 months

v2 <- greenland(9, 1/9, scale=12)
```

```
# treatment effect is expected to be between 1/4 and 4
# with prob 95% in 12 months
v3 <- greenland(4, 1/4, scale=12)

#Prior covariance matrix:
my.cov <- diag(c(v1,v2,v3))
for(i in 1:2)
    for(j in (i+1):3)
        my.cov[i,j] <- my.cov[j,i] <-
                    my.cor[i,j]*sqrt(my.cov[i,i])*sqrt(my.cov[j,j])

#Precision:
my.prec <- solve(my.cov)
```

The prior variances for $\boldsymbol{\beta}$ are determined as follows:

- The treatment effect $\exp(\beta_{\text{trt}})$ at baseline is expected to be close to unity, since the study was randomized. Specifically, we assume that the effect is between 9/10 and 10/9 with 95% probability, which corresponds to a prior variance of 0.0029 for β_{trt}.
- If the overall time trend would be zero, then we would expect the variation of the change of infection prevalence in the control group over the 12 months to be within the range $0.9 \rightarrow 0.1$ to $0.1 \rightarrow 0.9$ with 95% probability. This leads to a prior variance of 0.0087 for β_{time}, which we also consider as suitable for a non-zero overall time trend.
- The time-treatment interaction effect $\exp(\beta_{\text{time}\times\text{trt}})$ is expected to be between 1/4 and 4 with 95% prior probability after 12 months, which corresponds to a prior variance of 0.0035 for $\beta_{\text{time}\times\text{trt}}$.

As in the generalized g-prior, the covariances in $\boldsymbol{\Sigma}$ are determined based on the correlations in $\mathbf{X}^\top\mathbf{X}$ (the weight matrix $\mathbf{W}$ is the identity matrix here), where the design matrix $\mathbf{X}$ is based on two representative patients with regular visits, one from each treatment group:

$$\mathbf{X}^\top = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 2 & 3 & 6 & 9 & 12 & 0 & 1 & 2 & 3 & 6 & 9 & 12 \\ 0 & 1 & 2 & 3 & 6 & 9 & 12 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The resulting prior covariance matrix $\boldsymbol{\Sigma}$ is shown in Table 1.

	treatment	time	time $\times$ treatment
treatment	0.0029		
time	0.0027	0.0087	
time $\times$ treatment	0.0024	0.0039	0.0035

Table 1: Prior covariance matrix for toenail data example.

Besides of using a g-prior on the fixed effects $\boldsymbol{\beta}$ we also want to implement patient-specific random intercepts. This can be done in `pwGLM` by using the argument `FormExt` which adds a list of additional latent fields to the `r-inla` formula. The additional data, requested by the additional field defined by `FormExt`, must be handed over to `pwGLM` by the argument `AddDat`.

For the hyperparameters of the additional latent field, which is the variance of the random intercepts, we follow Fong et al. (2010), who have proposed an inverse gamma prior, as described in the main text.

```
#FormExt additional latent field f(...) for random intercepts.
#define inla-formula for additional latent field (called by pwGLMinla):
```

```
FormExttoe <- list(
                    "f(id1,
                       model=\"iid\",
                       hyper=list(theta=list(
                                        prior=\"loggamma\",
                                        fixed=F,
                                        param=c(0.5,  0.0164)))",
                       "id1")

#define AddDat (data.frame used in FormExt) called by pwGLMinla:
addtoeid <- data.frame(toe$id1)
names(addtoeid) <- c("id1")
```

Now we can call `pwGLM`, including additionally the arguments `FormExttoe` and `AddDat`:

```
#call pwGLM:
resultToe <- pwGLM(response=subset(toe, select=Response),
                    data=subset(toe, select=-Response),
                    fixedG=FALSE,
                    initialG=log(1),
                    priorG=prior.table,
                    priorMean=mean.toe,
                    priorPrecision=my.prec,
                    verbose=TRUE,
                    FormExt=FormExttoe, #adding random intercepts
```

```
        AddDat=addtoeid,  #patient-id used by FormExt
        Ntrials=Ntrials,  # number of trials for binomial
        intsep=TRUE)

#fixed effect estimates on OR scale:
exp(resultToe$betaMedian)
```

	treatment		**time**		**time** $\times$ **treatment**	
Prior	1.00	(0.90 to 1.11)	0.83	(0.69 to 1.00)	1.00	(0.89 to 1.12)
MLE	0.80	(0.40 to 1.59)	0.67	(0.64 to 0.70)	0.87	(0.81 to 0.93)
$g = 1$	0.93	(0.86 to 1.01)	0.69	(0.66 to 0.73)	0.90	(0.85 to 0.96)
hyper-g	0.92	(0.80 to 1.06)	0.69	(0.65 to 0.73)	0.89	(0.82 to 0.96)

Table 2: Prior, MLEs, and posterior median odds ratios with 95% CI for toenail data example.

Two models have been fitted with `R-INLA`, one with fixed $g = 1$, the other one with hyper-g prior (shown by the code above and stored as object `resultToe`). The posterior of g in the second model indicates no strong prior-data disagreement, with the weight of the prior reduced by a median factor of $\hat{g}$=2.3, but with considerable uncertainty (95%-CI: 0.67 to 10.1). The posterior odds ratio estimates, shown in Table 2, are therefore not very different with or without adaptive prior weighting. The ML estimates, obtained from an analysis with REML with the `R`-package `lme4`, show larger differences compared to the Bayesian analysis, in particular for the treatment effect β_{trt}.

References

De Backer, M., De Vroey, C., Lesaffre, E., Scheys, I., and De Keyser, P. (1998). Twelve weeks of continuous oral therapy for toenail onychomycosis caused by dermatophytes: A double-blind comparative trial of terbinafine 250 mg/day versus itraconazole 200 mg/day. *Journal of the American Academy of Dermatology* **38,** S57 – S63.

Fong, Y., Rue, H., and Wakefield, J. (2010). Bayesian inference for generalized linear mixed models. *Biostatistics* **11,** 397–412.

Held, L. and Sabanés Bové, D. (2014). *Applied Statistical Inference - Likelihood and Bayes*. Springer, Heidelberg.

Held, U., Sabanés Bové, D., Steurer, J., and Held, L. (2012). Validating and updating a risk model for pneumonia - a case study. *BMC Medical Research Methodology* **12,** 99.

Martins, T. G., Simpson, D., Lindgren, F., and Rue, H. (2013). Bayesian computing with INLA: new features. *Computational Statistics and Data Analysis* **67,** 68–83.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society - Series B* **71,** 319–392.

Sullivan, S. G. and Greenland, S. (2013). Bayesian regression in SAS software. *International Journal of Epidemiology* **42,** 308–317.